The Dissertation Committee for Victor Vaughan Symonds Certifies that this is the approved version of the following dissertation:

Genetic analyses of natural variation in the model plant *Arabidopsis thaliana*: neutral marker, quantitative genetic, and population genetic approaches

Committee:

Alan M. Lloyd, Supervisor

Donald A. Levin

Ulrich G. Mueller

Thomas E. Juenger

Robert K. Jansen

Genetic analyses of natural variation in the model plant *Arabidopsis thaliana*: neutral marker, quantitative genetic, and population genetic approaches

by

Victor Vaughan Symonds, B.A., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December, 2004

Acknowledgements

Throughout my graduate career I have benefited from the wisdom and kindness of others – too many to mention, in fact. I must, however, thank my committee members for their perennial encouragement and generosity of time. Leading this list, of course, is my advisor, Alan Lloyd, who provided an environment in which any reasonable, and sometimes unreasonable, path could be followed . . . and usually was. Finally, Jennifer Tate provided daily personal and professional support for which I am eternally grateful.

Genetic analyses of natural variation in the model plant *Arabidopsis thaliana*: neutral marker, quantitative genetic, and population genetic approaches

Publication No._____

Victor Vaughan Symonds, Ph.D.

The University of Texas at Austin, 2004

Supervisor: Alan M. Lloyd

The model flowering plant *Arabidopsis thaliana* is a popular system for studies of development, molecular biology, physiology, and more recently, ecology and evolution. The work described here represents several approaches toward understanding natural variation within this species. A molecular marker approach was taken to evaluate the distribution of neutral genetic variation. This study examined variation at 20 microsatellite loci in 126 accessions of *A. thaliana*. Substantial variability in mutation pattern was found among loci, most of which is not explained by the assumptions of traditional mutation models. Instead, it was discovered that the degree of locus diversity is strongly correlated with the average number of contiguous repeats within a locus, supporting a strong role for repeat disruptions in stabilizing microsatellite loci by reducing the substrate for polymerase slippage and recombination. Cluster analyses generated from these same data demonstrate the potential of microsatellite loci for resolving relationships among accessions of *A. thaliana*. To explore the genetic bases of

phenotypic variation observed within *A. thaliana*, a quantitative trait locus (QTL) mapping approach was employed focusing on a broadly distributed, selectively important trait, trichome density. While such experiments are typically conducted in a single mapping population, to better characterize the genetic architecture of this trait, QTL were mapped in four recombinant inbred line populations. The results of this work identify eight new QTL for trichome density and show that many lineage-specific alleles that either increase or decrease trichome density persist in natural populations. Several of the QTL identified in these studies mapped near a candidate gene. One of these loci, *ATMYC1*, was particularly intriguing because previously it had not been linked to the trichome initiation genetic pathway. To test the hypothesis that *ATMYC1* is a QTL for trichome density, genetic complementation tests and a screen for molecular variation were employed. Combined, these data revealed that alleles of *ATMYC1* from different accessions of *A. thaliana* vary in their abilities to recover the mutant *atmyc1* trichome phenotype, indicating that *ATMYC1* is a QTL for trichome density. Sequence analyses of alleles of *ATMYC1* revealed high levels of molecular divergence, suggestive of strong intraspecific, divergent selection.

Table of Contents

List of Tables

List of Figures

Chapter1: Introduction

This chapter is intended as an overall introduction to the dissertation. All of the work described herein was born out of very broad interests in biological diversity, the mechanisms by which it is generated, and the forces that dictate its maintenance and loss. As these interests span from molecular to phenotypic diversity, all of the work described here was conducted using the model flowering plant, *Arabidopsis thaliana* (Cruciferae). Like many model organisms, *A. thaliana* has suffered a reputation as a "lab rat" that is only somewhat accurate. The reality is that natural phenotypic variation within this predominantly selfing species is comparable to levels seen within many genera. Natural populations of *A. thaliana* are distributed across the northern hemisphere from the northern tip of Africa to the Arctic Circle. Seed stocks have been derived from populations from around the world and are referred to as stocks, accessions, lines, strains, or ecotypes. For the sake of clarity, the term accession is generally used throughout the following chapters; accessions refer to specific seed stock numbers that are less likely to be confused than ecotypic names. The work described in the next three chapters utilized this readily available source of natural variation to address specific molecular, quantitative, and population-genetic questions.

Chapter two describes a survey of neutral marker diversity within *A. thaliana* and a mechanism thought to regulate microsatellite mutation frequency. Microsatellite loci have become a preferred marker type because they often reveal molecular variation in systems that have otherwise shown none. To assess variation among microsatellite loci

for gene diversity, 126 accessions of *A. thaliana* were screened at twenty microsatellite loci. Goodness of fit tests for several common microsatellite mutation models typically failed to explain the diversity observed for loci of *A. thaliana*. This lead to an investigation of the potential role for repeat interruptions in stabilizing microsatellite alleles using six markers that were representative of different modes of gene diversity (very high and very low diversity). Several alleles of each of locus were cloned and sequenced. Sequence data revealed the presence of several repeat interruptions within the most frequent alleles sampled from low diversity loci, whereas, among high diversity loci, either no such interruptions were present or they only appeared near extreme ends of the repeat region, leaving long tracks of perfect repeats. It turns out that the number of perfect repeats possessed by the average or most frequent alleles are strongly correlated with gene diversity for a locus. These data show that the level of polymorphism detected for a given microsatellite locus is dependent upon, and perhaps more importantly, is predicted by allele length and repeat interruptions. Beyond a better characterization of neutral marker diversity within *A. thaliana*, these findings have important general implications for 1) understanding the sources of microsatellite marker diversity and 2) inferences drawn from microsatellite locus variation. This second chapter has been published in *Genetics* (SYMONDS and LLOYD 2003).

The third chapter characterizes the genetic architecture of a phenotype known to be of selective importance in wild populations of *A. thaliana*, trichome density. Trichomes are epidermal cells that are known to play important protective roles against phytophagous insects, UV radiation, and desiccation. Among accessions of *A. thaliana*, there is tremendous diversity for trichome density, ranging from completely glabrous

plants to very densely pubescent. To better understand the genetics that underlie this variation, quantitative trait loci (QTL) were mapped in four recombinant inbred line populations made from six accessions of *A. thaliana*. Because three populations have a parent in common, this set of experiments has the added benefit of a test-cross design. Mapping results identified nine QTL located throughout the genome. This is in sharp contrast with the only previous QTL study of trichome density, which identified a single QTL with very large effect (LARKIN et al. 1996). The Larkin et al. study has been cited as evidence for a single gene of large effect controlling natural phenotypic variation. Contrary to this, our data, which were derived from a broader sampling of accessions, indicate that natural allelic variation at several loci contributes to observed trichome density variation within the species. Although the mapping of QTL provides a means by which one may estimate several genetic parameters that affect trait variation, mapping results should be considered a set of testable hypotheses, rather than an endpoint of inquiry; chapter four describes work that tests one of these hypotheses. The third chapter is in press at *Genetics* (SYMONDS et al. in press).

The work described in the fourth chapter was born out of the quantitative genetic analyses reported in chapter three. Among the QTL identified in chapter three, several of them mapped near good candidate loci – genes with known mutant trichome phenotypes. A more intriguing candidate than these, however, was a gene with no known trichome phenotype, but that is a paralog of known trichome initiation regulators; this gene is *ATMYC1*. Two approaches were taken to examine whether *ATMYC1* is a QTL for trichome density: genetic complementation tests and a screen for allelic molecular variation. Complementation tests included a traditional crossing scheme and transgenics.

Both tests showed that alleles of *ATMYC1* cloned from different accessions do not recover the mutant *atmyc1* trichome phenotype equally well. Specifically, the L*er*-0 allele behaves like a nonfunctioning null, unable to increase trichome number (surrogate for trichome density) in an *atmyc1* mutant background. By contrast, alleles cloned from three other accessions, Col-0, Sha, and CVI, all restored the trichome phenotype back to wildtype levels. Combined, these data indicate that *ATMYC1* is a QTL for trichome density. Sequence variation at *ATMYC1* revealed two well-supported clusters of alleles and several subtypes within each of the major clusters. Furthermore, it appears that subtypic classification may predict relative allele function, a testable hypothesis that will be pursued in the future. Work reported in this last chapter will be submitted for publication in the near future.

Chapter 2: An analysis of microsatellite loci in *Arabidopsis thaliana*: mutational dynamics and application

ABSTRACT

Microsatellite loci are among the most commonly used molecular markers. These loci typically exhibit variation for allele frequency distribution within a species. However, the factors contributing to this variation are not well-understood. To expand on the current knowledge of microsatellite evolution, 20 microsatellite loci were examined for 126 accessions of the flowering plant, *Arabidopsis thaliana*. Substantial variability in mutation pattern among loci was found, most of which cannot be explained by the assumptions of the traditional Stepwise Mutation Model or Infinite Alleles Model. Here it is shown that the degree of locus diversity is strongly correlated with the number of contiguous repeats, more so than the total number of repeats. These findings support a strong role for repeat disruptions in stabilizing microsatellite loci by reducing the substrate for polymerase slippage and recombination. Results of cluster analyses are also presented, demonstrating the potential of microsatellite loci for resolving relationships among accessions of *A. thaliana*.

INTRODUCTION

Microsatellite loci are tandemly repeated DNA motifs of 1-6 bp in length; they are also referred to as simple sequence length polymorphisms (SSLPs), simple sequence repeats (SSRs), simple tandem repeats (STRs), and variable number tandem repeats (VNTRs). These loci occur at high frequency in all eukaryotes examined (KATTI et al.

2001) and at some lower frequency in prokaryotic genomes (METZGAR et al. 2001). The use of microsatellite loci as polymorphic DNA markers has expanded considerably over the past decade both in the number of studies (ESTOUP and ANGERS 1998) and in the number of organisms (BARKER 2002), primarily due to their facility and power for population genetic analyses. Microsatellite loci are typically highly variable, even in organisms that otherwise display little genetic variation (ZWETTLER et al. 2002), are relatively straightforward to identify (ZANE et al. 2002), and can be scored via many different methods. Although originally described from humans for use in genetic fingerprinting (LITT and LUTY 1989), microsatellite locus use today includes genetic mapping (e.g., MCCOUCH et al. 1997; SAKAMOTO and OKAMOTO 2000), assessments of genetic diversity (CRUZAN 1998; DRISCOLL et al. 2002), forensics (GILL et al. 1985; KUBO et al. 2002) and studies of human genetic disease proliferation (CUNNIFF 2001; RANUM and DAY 2002).

Microsatellite loci increase and decrease in length due to polymerase slippage during DNA replication (ECKERT et al. 2002) and recombination (RICHARD and PAQUES 2000), both of which are consequences of having a series of identical tandemly repeated units. With these phenomena in mind, discussions of microsatellite evolution primarily center around two models, the stepwise mutation model (SMM) and the infinite alleles model (IAM) (BALLOUX and LUGON-MOULIN 2002). In short, the SMM suggests that the mutation of microsatellite alleles occurs by the loss or gain of a single tandem repeat, and the IAM describes mutations involving the loss or gain of any number of repeats, but always generates new, previously unsampled alleles (see review by ESTOUP and CORNUET 1999). One commonality between these two models is that they only consider

changes in tandem repeat number. More recently it has been suggested that microsatellite locus evolution is most strongly influenced by the balance between locus length and point mutation rate (KRUGLYAK et al. 1998). Specifically, longer microsatellite alleles are hypothesized to be more prone to generate new length variants than are shorter alleles (WIERDL et al. 1997). However, nonrepeat mutations (substitutions, insertions, and deletions) that interrupt perfect tandem repeats affect the function of length. The disruption of a set of tandem repeats by any process, including indels and point mutations, in effect, lessens the number of perfectly repeated units, and is expected to reduce the likelihood of locus evolution (ROLFSMEIER and LAHUE 2000). Despite advances in development of molecular evolution models and the wide-spread use of microsatellite markers, detailed analysis of microsatellite evolution and the underlying forces remain limited to relatively few studies representing even fewer organisms (for examples see: NOOR et al. 2001; VIGOUROUX et al. 2002).

Arabidopsis thaliana has long been a model genetic and molecular system for plant biology. Recently, natural variation within this species has come into focus (ALONSO-BLANCO and KOORNNEEF 2000), expanding its utility toward addressing evolutionary and population biology questions. Unfortunately, the genetic infrastructure, including mapping data, in place for the few most-commonly used accessions of *A. thaliana*, does not yet extend to the several hundred wild-collected accessions available. An examination of microsatellite variation within *A. thaliana*, therefore, serves at least two purposes: improving upon the genetic tools available for this model organism and expanding our knowledge of microsatellite evolution.

Previous studies on *A. thaliana* microsatellite loci have shown that they are abundant (CASACUBERTA et al. 2000; KATTI et al. 2001) and highly variable (INNAN et al. 1997; VAN TREUREN et al. 1997; CLAUSS et al. 2002). However, these works are limited to fewer than 50 accessions and the studies minimally overlap in marker usage. To develop the utility of microsatellite loci among wild accessions and to investigate factors affecting mutation patterns at these loci, we have gathered size and sequence data for a diverse collection of *A. thaliana* accessions. We find substantial variability in mutation pattern among microsatellite loci and among accessions, most of which specifically conforms to neither the SMM nor the IAM. Contributing to this variation is sequence complexity and the presence of repeat disruptions within loci. Here we show that high diversity loci tend to possess long stretches of contiguous repeats, while low diversity loci are either uninterrupted with few total repeats or contain repeat interruptions that result in few contiguous repeats. Further, sequence data indicate that there is a wealth of intraspecific, potentially phylogenetically informative variation at these loci, an important point in a model system for which we possess little genealogical information.

MATERIALS AND METHODS

Plant materials

Genetic variation among 120 "wild" accessions and several commonly used reference accessions (including Col-0, L*er*, and WS) of *Arabidopsis thaliana* was surveyed (TABLE 2.1). Line selection was based on global population coverage and, for a

subset of lines, local proximity. That is, a few nested accessions including separate collections made from near the same location were selected. Although microsatellite size data exist for the three reference accessions, different size scoring methods tend to yield varying results (personal observation). Therefore, the reference accessions were included in our analyses in order to derive data directly comparable with all other accessions included. Two stocks, Cal-0 and Tac-0, were generously provided by Johanna Schmitt and Lisa Dorn. Three of the reference accessions used were lab stocks. All remaining seed stocks were acquired from the *Arabidopsis* Biological Resource Center. Although all accessions of *A. thaliana* are reportedly nearly completely homozygous (BERGELSON et al. 1998), all accessions included here underwent at least one round of additional selfing in our lab prior to genotyping. Seed for all lines were imbibed in water and vernalized at 4° for three days prior to germination at 22° under 24 hours light.

Microsatellite survey

Total DNAs were extracted from several rosette leaves of a single individual for each accession following a modified CTAB method (modified from DOYLE and DOYLE 1987). Approximately 50 ng of total DNA were used as template in individual microsatellite amplification reactions.

All lines were screened at 20 microsatellite marker loci. These loci were selected to provide approximately equal coverage across the genome at a density equivalent to that required for rough-scale mapping (~ every 30 cM), taking into consideration both the distance between pairs of markers and between centromere and chromosome end positions (TABLE 2.2). Primer sequences for all loci were acquired through the

Arabidopsis Information Resource (TAIR, http://www.arabidopsis.org), all of which were originally described by BELL and ECKER (1994) and LUKOWITZ et al. (2000). Each locus was amplified by PCR and fluorescently labeled by one of two methods; either the forward primer in each reaction was labeled directly with one of the three dyes (D2, D3, and D4) used on Beckman-Coulter instruments or an M13 tailing scheme was followed as described by BOUTIN-GANACHE et al. (2001), whereby the forward primer was 5'-tailed with the M13 forward sequence and used in conjunction with a 15-fold excess of a fluorescently labeled M13 forward primer. All primers used in amplification reactions were synthesized by ResGen (Invitrogen, Inc.). The switch was made to the M13 tailing scheme because it requires only three fluorescently labeled primers, rather than independently labeling all forward primers.

Amplification reactions were carried out in 10 μL volumes containing 1X PCR buffer (Invitrogen, Inc.), 1.5 mM $MgCl_2$, 50 μM each dNTP and either 600 nM each primer (for reactions with labeled forward primer) or 150 nM labeled M13 and reverse primers and 10 nM unlabeled forward primer (for M13 tailed primer scheme). Approximately 50 ng of each DNA extraction was used as template for individual locus amplification in a standard 96-well plate format. Standard amplification conditions consisted of 95° for 3 min, 30 cycles of denaturing at 94° for 1 min, annealing at 55° for 1 min, and polymerization at 72° for 1 min, followed by a final extension for 6 min at 72°. As the annealing temperature for the M13 primers is lower than that of the average SSLP primer, amplification conditions for the M13 scheme were modified by lowering the annealing temperature to 52°. Although two amplification schemes were used, the

amplification conditions for each locus was consistent among all accession templates amplified and no significant difference in amplification rate was observed between the two protocols.

Microsatellite length polymorphisms were detected and scored by capillary electrophoresis on a Beckman-Coulter CEQ 2000XL DNA Analyzer. Although all amplification reactions were carried out individually, the use of three different dyes allowed for the pool-plexing of samples during separation and allele sizing. Typically, the PCR products of three separate reactions for one individual, each labeled with a different dye (D2, D3, and D4) were pooled. The pooled products were then purified in vacuum filter plates (Millipore MANU030) at 20 inches Hg for four minutes (manufacturer's specifications), and subsequently eluted in 30 µL H_2O. 1.25 µL of each cleaned, pooled sample was then added to 0.5 µL of 400 bp size standard (labeled with D1 dye) and 38 µL of Sample Loading Solution (Beckman-Coulter) in a well of a 96-well sample plate, and overlaid with mineral oil. Each pool-plexed sample was separated on the CEQ using the standard Frag-1 method. This pool-plexing system resulted in the separation of products at three different loci simultaneously through a single capillary along with an internal size standard. Fragments were sized using the default fragment analysis protocol for the appropriate set of dyes used (AE2 or PA1 options).

Microsatellite Data Analyses

The CEQ raw data from each run were analyzed using the appropriate dye mobility calibration settings for each dye and the default Fragment Analysis settings for the 400 bp size standard. Alleles reported here reflect the amplification product size, as

scored on a CEQ 2000XL DNA analyzer. Simply inferring the number of repeats from size data ignores potentially informative data from indels. Often alleles are sized based on assumptions regarding the locus, for example, alleles at a dinucleotide repeat locus are often assumed to only fall into size classes two bp apart, however, our sequence data show real indels and real one bp differences among alleles at dinucleotide repeat loci in our data set. Therefore, we report all observed size classes, regardless of the repeat type.

Microsatellite cloning and sequencing

To investigate the nature of length variations within loci, several alleles were cloned and sequenced for six loci. Three loci were randomly selected from among the low diversity loci (nga1107, nga1145, nga129) and three from among the high diversity loci (CIW7, nga172, and nga8). Individual alleles were amplified with unlabeled (no dye) forward and reverse primers as described in the preceding section from individual accessions. One microliter of PCR product was then added to a cloning reaction using the TOPO-TA cloning kit (Invitrogen, Inc.). Colonies with inserts were initially identified by blue/white screening, followed by PCR amplification from individual colonies, and size confirmation on agarose gels. Multiple clones for each reaction were identified and plasmid DNA mini-preparations were prepared from selective overnight liquid cultures. DNA mini-preparations were carried out following a modified SDS protocol where DNA precipitation is preceded by separate phenol and chloroform extractions. Approximately 500 ng of vector with insert were used as template in sequencing reactions using either the T7 or M13-reverse primer. Sequencing reactions were purified using Sephadex G-50 columns and the sequences were analyzed on an MJ

BaseStation DNA analyzer. Post-run data were processed using the Cartographer v. 1.2.4sg software (MJ Research, Inc.). Sequence alignments for alleles of each locus were carried out using Megalign (DNAStar, Inc.).

Associations between locus length and locus diversity

Associations between the genetic diversity of a locus and some measure of locus length, typically mean length, is commonly reported for microsatellite loci (BACHTROG et al. 2000; MORIGUCHI et al. 2003). To investigate this association for the loci examined here, the mean allele size was determined for each locus. That allele, or the nearest in size, was cloned and sequenced from multiple accessions for 10 loci, as described above. From these sequences and available Col sequence, the total number of repeats was counted or inferred for each locus by subtracting the shared, non-repeat flanking sequence from the total locus length. Association strength between repeat number and locus diversity was assessed by calculating the correlation coefficients between the two; both Pearson product moment correlation and Spearman's coefficient of rank correlation were calculated. To examine the potential role of repeat interruptions on locus diversity, from those same sequences the largest number of contiguous repeats was counted. For example, in the following sequence, ACTGAGAGATTGAGAGAGACTT, the total number of repeats is seven, and the largest number of contiguous repeats is four. Again, association strength was determined by calculating correlation coefficients between the largest number of contiguous repeats and locus diversity. Because different repeat types often have different mutation rates (BACHTROG et al. 2000; HILE et al. 2000), for these analyses data were partitioned into two groups, according to repeat type:

15 “GA” repeat loci and four “TA” repeat loci; the one trinucleotide repeat locus in this study was omitted from these analyses. To further examine these relationships, data for the GA repeat loci were divided into groups of high and low locus diversity.

Genetic analyses

Gene diversity estimates for each locus were calculated by:
$n(1-\Sigma p_i^2)/(n-1)$, where n is the number of samples and p_i is the frequency of the i^{th} allele, following the methods of NEI (1973)and MATSUOKA et al. (2002). The value n is used here in place of $2n$ because all *A. thaliana* accessions are expected to be near completely homozygous due to inbreeding.

The fit of each locus’ distribution to expected distributions under three different mutation models, the Stepwise Mutation Model (SMM), the Infinite Alleles Model (IAM), and an intermediate Two-Phase Model (TPM) was tested using the program BOTTLENECK (CORNUET and LUIKART 1996). Because of sampling, data for all accessions were treated as a single population, which is not ideal, but unavoidable. Observed allele frequencies and sample sizes were input parameters. These analyses provide a test statistic for the probability that an observed allele distribution with a given heterozygosity (gene diversity) was generated under each of the three mutation models.

To describe the distribution of alleles for each locus, measures of skewness (g_1) and kurtosis (g_2) were calculated following SOKAL and ROHLF (1995). Significant differences between low and high diversity loci were tested for by a simple *t*-test for each measure. Because of different mutation rates between dinucleotide and trinucleotide loci

(CHAKRABORTY et al. 1997; SIA et al. 1997), the GapAB locus was omitted from these analyses.

For similarity analyses, allele size class data were transformed into alphanumeric codes. From this transformed data set, pairwise distances were obtained based on the proportion of shared alleles, as implemented in PAUP*4.0b10 (SWOFFORD 2002). As the complete evolutionary history of *A. thaliana* accessions is partially reticulate and therefore cannot be accurately represented by a bifurcating tree, a majority-rule (70%) consensus tree of 1000 independent cluster analyses using unweighted pair group method using arithmetic averages (UPGMA) is presented to simply illustrate genetic similarity among accessions. In the course of building trees, cluster analyses have to randomly break ties between equivalent relationships. As a result, there is a stochastic component to resulting trees. One thousand independent UPGMA analyses were run on the complete data set and only relationships consistent with the 70% majority rule are presented to provide a more rigorous analysis and conservative tree. More detailed analyses aimed at reconstructing the intraspecific phylogeny of *A. thaliana* will be presented elsewhere.

Because low diversity and high diversity loci may be influenced by differing mutation dynamics, we conducted a partition homogeneity test implemented in PAUP (FARRIS et al. 1994; FARRIS et al. 1995; SWOFFORD 2002), which tests for the probability of significant conflict between data partitions with regard to phylogeny. The total dataset was partitioned into two mutually exclusive groups, conservatively excluding the nga129 locus altogether because of its intermediate diversity measure. The low diversity group included all loci with gene diversity measures less than 0.70 and the high diversity partition included all loci with gene diversity measures above 0.80.

RESULTS

Amplification fidelity

Amplification success varied both across the 20 loci and among the 126 accessions of *Arabidopsis thaliana*. Amplification frequencies for each of the twenty loci investigated are listed in TABLE 2.2. Amplification success ranged from 77 to 98% across loci and from 70 to 100% among accessions (excluding four accessions; data not shown), with a total of 90% amplification success. No significant correlation was found between amplification success and any measure of locus diversity (analyses not shown). Of the 2,526 marker-by-individual data points, only four (0.2%) were found to be heterozygous. This frequency is similar to that reported for 12 accessions of *A. thaliana* by CLAUSS et al. (2002). Amplification of two loci, GapAB and CIW10 consistently yielded two products for all accessions. In each case, the size of one of the products was constant among all accessions and the other varied. Considering the high level of gene duplication within *A. thaliana* (VISION et al. 2000), this observation may represent the simultaneous amplification of two distinct loci. In each case, only the variable allele was included in our analyses.

Allelic diversity within and among loci

There is a high degree of variation among microsatellite loci for allelic diversity (FIGURE 2.1). The most striking differences are in the variation at a locus (number of alleles scored) and how that variation is distributed among alleles at a locus (gene diversity). These two measures are reported for all loci in TABLE 2.2. The average

number of alleles detected per locus is 17.6 (range = 4-38). The average gene diversity estimate from our data is 0.76 (range = 0.41-0.96; FIGURE 2.2), and does not differ appreciably from that of 0.79, reported by INNAN et al. (1997). Several different distribution patterns of allelic diversity are evident (FIGURE 2.1). For further analysis, loci were split into two very broad categories (FIGURE 2.2): high diversity (above the mean) and low diversity (below the mean).

High diversity loci tend to be either somewhat normally distributed or strongly positively skewed ($\overline{X}$ skewness = 1.11). These loci also tend to have leptokurtotic distributions ($\overline{X}$ kurtosis = 1.71). Low diversity loci show similar distribution patterns to the high diversity loci, typically positively skewed ($\overline{X}$ skewness = 1.79) and leptokurtotic ($\overline{X}$ kurtosis = 3.98), but to a significantly greater degree ($p<0.05$ for both tests). The tendency of microsatellite loci to mutate to larger allele sizes more frequently than to smaller sizes (becoming positively skewed) is well documented (RUBINSZTEIN et al. 1999; BROHEDE et al. 2002).

Sequence results

As initially scored, PCR products for many loci displayed single basepair differences among alleles, however, our sequence data showed that ~95% of single basepair differences initially detected were artifactual. Re-examination of the original electropherograms determined that these discrepancies were attributable to the inconsistent nontemplate-dependent terminal transferase activity of *Taq* polymerase that adds a single deoxyadenosine (A) to the 3' ends of PCR products. Though at a low frequency, instances of true single basepair differences were also revealed (e.g., see

alleles of locus nga129 in FIGURE 2.3). All sequenced size outliers proved to be the expected locus.

Molecular variation at high diversity loci

Individual alleles of three loci demonstrating high gene diversity were cloned and sequenced. An allelic alignment for a representative locus is shown in FIGURE 2.3 (Nga8). All 34 alleles sequenced from these loci were found to be either "perfect", that is, without interruptions of any kind within the repeated region (Nga172 and CIW7), or possessing nearly fixed interruptions in the extreme end of the repeat region (Nga8). With this one exception, the only source of size variation identified at these high diversity loci was changes in repeat number. Although point mutations were identified in flanking regions, no insertions and deletions were revealed.

Molecular variation at low diversity loci

The three low diversity loci for which alleles were sequenced, each revealed alleles with interruptions within the repeated region (FIGURE 2.3). In each case, the interruptions consisted of two basepair insertions, back-to-back nucleotide substitutions, or some combination thereof; the origins of interruptions within tandemly repeated regions typically cannot be distinguished from among these possibilities.

Of the seventeen alleles of the nga129 locus that were sequenced, only one size class (the most common) revealed an interruption. The 190 bp allele possesses a "CT" doublet within the repeat region, along with a two bp mutation, that immediately flanks the 3' end of the microsatellite locus. These two mutations were always found to be

linked. That is, no alleles were sequenced that possess one mutation and not the other. This pair of mutations was only found in the 190 bp allele, and all eight alleles of this size that were sequenced are identical. The remaining variation detected among alleles at this locus appears to be the result of varying repeat number only.

Upon sequencing 25 alleles from the nga1145 locus, an "AA" interruption three repeat units from the 3' end of the locus was discovered. Unlike the nga129 locus, this interruption is evident in many alleles (size classes), rather than only the most common allele. Other sources of variation at this locus include a unique "GG" mutation, immediately flanking the AA interruption, and a single finding of an apparent duplication event comprised of the entire microsatellite locus (Accession # 6672). For this locus also, all remaining allelic diversity appears to be due to repeat number variation.

As with most loci, the primary source of size variation is change in repeat number for the nga1107 locus also, however, this locus is the most complex with regard to interruptions. It consists of four "GA" repeat regions, separated by 11 bp, 14 bp, and two bp interruptions, from the 5' to 3' ends, respectively. The second of these three interruptions appears to be a complex of successive variable number tandem repeat (VNTR) loci (GCGC/TT/AAA/CCC/TA). Excepting its absence from one line, however, no sequence variation was uncovered within this complex among the seven accessions sequenced. Interestingly, the accession missing this insertion is the common reference strain, Col-0. Col-0 also lacks the second insertion and, despite these deletions (or lack of insertions), possesses the longest allele sampled at this locus due to many more repeats.

Relationship between contiguous repeat length and gene diversity for all loci

Correlation analyses show a general positive relationship between number of repeats possessed by the mean allele of a locus and locus diversity, however, this relationship varies depending on how the data are partitioned (Table 2.3). For the comparisons made, Pearson's and Spearman's correlation coefficients are in general agreement, therefore, unless stated otherwise, discussion applies to results of both tests. For the 15 loci with GA repeats, the total number of repeats possessed by the mean allele does positively correlate with locus diversity. However, the number of uninterrupted repeats demonstrates a stronger and (for Pearson's) more significant correlation with locus diversity. Analyses including only high or low diversity loci show the same trend, with one clear exception; for low diversity loci, the total number of repeats in the mean allele shows *no* significant correlation with locus diversity. The four TA repeat loci show the same general positive correlation between locus diversity and repeat number, again, with the number of contiguous repeats being more tightly correlated with diversity than the total number of repeats. The smaller sample size for TA repeat loci precluded more detailed analyses.

Testing the SMM, TPM, and IAM

Results of mutation model tests are shown in TABLE 2.2. Of the twenty loci examined, five potentially fit all three models of evolution tested and six display distributions that do not differ significantly from the expected distribution under any of the three models tested (SMM, TPM, and IAM). Only three loci rejected two of the three models, suggesting the third as a reasonable fit. On average, low diversity loci show a

much higher model rejection rate than do high diversity loci (Table 2.4), although the relative rejection rate among tests is consistent between the two sets of loci. Consistent with other reports (see review by ELLEGREN 2000), the SMM was the most frequently rejected model (13/20 loci), although most loci show hallmarks of SMM-like evolution (FIGURE 2.1). Under the SMM and TPM, there is a general heterozygosity deficiency (19/20 and 17/20 loci, respectively), and under the IAM, there are an equal number of loci with heterozygosity excess and heterozygosity deficit.

Performance of microsatellite data in cluster analyses

To evaluate the performance of *A. thaliana* microsatellite loci for estimating intraspecific relationships, a majority-rule consensus tree based on 1000 UPGMA cluster analyses was generated (FIGURE 2.4). Because of the inclusion of particular pairs and sets of accessions, many relationships could be predicted. For example, groups of accessions collected from the same locale were included (e.g., Nok-0. Nok-1, Nok-2, etc.) and were expected to cluster together. The cluster analysis presented here reveals many groupings that are consistent with predicted associations. A selection of expected clusters are highlighted in FIGURE 2.4 and are discussed below. Interestingly, partition homogeneity tests revealed no significant incongruence between low and high diversity loci (P=0.90).

DISCUSSION

Over the past decade, the frequency of microsatellite locus use has increased considerably (see reviews by ESTOUP and ANGERS 1998; ELLEGREN 2000). Despite this

newfound popularity, detailed examinations of microsatellite mutation patterns, and the forces that generate and maintain microsatellite diversity remain restricted to relatively few organisms. Here we present analyses of allelic variation at 20 loci for 126 accessions of *Arabidopsis thaliana*. We provide sequence data supporting a strong role for repeat interruptions that result in relatively short repeat segments in stabilizing microsatellite loci, and present a cluster analysis of all accessions.

Amplification fidelity

Each primer pair successfully primed amplification in an average of 90% of all accessions examined. This amplification rate is similar to that reported in other studies of *A. thaliana* microsatellites (CLAUSS et al. 2002; INNAN et al. 1997). The consistency of amplification rate among reports, and repeated attempts at amplifying individual null alleles (no amplification product) would seem to argue that the remaining null alleles are mainly due to sequence divergence within priming sites or deleted loci, rather than spurious amplification failure. Furthermore, accessions suspected to be closely related show similar null allele patterns (not shown).

Forces affecting mutation patterns

As has been reported in other systems (BACHTROG et al. 2000; BROHEDE et al. 2002), several different patterns of allelic distribution were revealed among the microsatellite loci of *A. thaliana*. The mutation models typically invoked to explain microsatellite distribution patterns are the Stepwise Mutation Model (SMM), the Infinite Alleles Model (IAM) or some combination thereof (e.g., the Two-Phase Model (TPM)).

However, observed distribution patterns rarely fit the stringent SMM (ELLEGREN 2000; SHRIVER et al. 1993) and empirical evidence documenting independent identical mutations argue against the IAM (BROHEDE et al. 2002; THUILLET et al. 2002). Indeed, more than one half of the loci examined here have distributions that either differ significantly from and thus reject all models, or fit all models equally well (TABLE 2.2), effectively supporting none. An alternative to simple mutation dynamics in explaining the observed model-fit results is that some aspect of population demography has resulted in the observed allele distributions. Two of the mutation models support this alternative, showing strong trends toward heterozygosity deficit (17/20 loci for the TPM and 19/20 for the SMM), a finding that is consistent with hypotheses regarding the relatively recent and rapid expansion of *A. thaliana* global populations (Sharbel et al. 2000), while only under the IAM is the assumption of equilibrium met. Unfortunately, both mutation dynamic and demographic interpretations are compromised by violations of certain test assumptions, specifically that the sample represents a single contiguous population at mutation-drift equilibrium.

Beyond these models, it has been suggested that microsatellite locus equilibrium is a balance between polymerase slippage rate and mutation rate (KRUGLYAK et al. 1998; SCHUG et al. 1998). In short, the longer the string of uninterrupted repeats (e.g., AGAGAGAG), the more likely is the generation of new alleles via slippage and recombination. Any mutation within the repeated region that causes an interruption (e.g., AGAG***TTT***AGAG) will effectively split the original repeat region into two shorter segments. This is expected to increase locus stability (i.e. reduce the generation of new alleles), simply by reducing the substrate for polymerase slippage and recombination.

This model would seem to explain our typical finding of repeat interruptions in low diversity loci and longer stretches of uninterrupted repeats within high diversity loci.

To investigate this further, we examined the relationship between mean allele length and locus diversity for different data partitions (Table 2.3). If repeat disruptions stabilize loci simply by breaking them into smaller segments, then the degree of stability conferred should be dependent upon the lengths of the resulting repeat segments. This was tested by comparing the strengths of association between locus diversity and (1) the total number of repeats possessed by the mean allele at a locus and (2) the largest number of contiguous repeats possessed by the mean allele. The results show that gene diversity is more strongly correlated with the number of contiguous repeats than it is to the total number of repeats (Table 2.3); the number of contiguous repeats accounts for 12% (all GA repeats), 66% (low diversity GA repeats) and 40% (TA repeats) more of the observed variation in genetic diversity, as determined by comparing coefficients of determination (r^2). The nature of this difference becomes evident when high diversity and low diversity loci are examined separately; this was only possible for the GA repeat loci, where sample size was sufficient. The correlation with diversity turns out to be identical for total repeat number and contiguous repeat number for high diversity loci. This is a result of high diversity loci tending not to be interrupted, which means that the total number of repeats is equal to the contiguous number of repeats. Conversely, low diversity loci demonstrate *no* (Spearman's) and very weak (Pearson's) relationships between total number of repeats and diversity (Table 2.3), whereas including only the number of contiguous repeats yielded some of the strongest associations with diversity observed among all complete and partitioned datasets. This provides strong evidence supporting a role for repeat

disruptions in locus stability, one that is highly dependent upon placement of the interruption and the lengths of the remaining contiguous repeats. Because several of the low diversity loci are without interruptions, this tight relationship also indicates that interrupted loci with few contiguous repeats behave in a manner similar to uninterrupted loci with few total repeats. As marker selection is often governed by criteria such as gene diversity, contiguous repeat number for mean allele size may provide a valuable predictor of marker utility. How broadly this relationship holds will require similar analyses in other organisms.

Size homoplasy

At any taxonomic level, the issue of size homoplasy in microsatellite data sets is an important and complicated one (ESTOUP et al. 2002). Size homoplasy can arise in a number of ways. Given the high mutation rate estimates for microsatellite loci (HANCOCK 1999), convergence on repeat number via slippage is likely the most common type, and unfortunately, impossible to detect *a posteriori*. As such, its frequency in *A. thaliana* cannot be addressed in our analysis. Another type of homoplasy involves mutations within the microsatellite locus other than changes in repeat number that result in size convergence. Through sequencing we have detected non-slippage mutations (e.g., point mutations and insertions) within repeat regions that have led to size homoplasy. Each of the low diversity loci possessed two bp repeat interruptions that could easily be misinterpreted as repeat number variation from size data alone (see FIGURE 2.3). These findings suggest a strong potential for this type of size homoplasy. Fortunately, these cases are easily detected via sequencing. A third homoplasy type for microsatellite loci

involves DNA insertions and deletions flanking the repeat region (GRIMALDI and CROUAU-ROY 1997). Our sequence data revealed predominantly point mutations in the immediate flanking regions; no insertions or deletions were discovered in 84 sequenced alleles (data not shown). Interestingly, a large-scale analysis of microsatellite marker loci in maize (MATSUOKA et al. 2002a) has shown that the most common source of variation is indels flanking the repeat locus. This sharp contrast in intraspecific sources of size variation underscores the need for more detailed microsatellite studies.

Cluster analyses

Previous efforts toward genealogy reconstruction within *A. thaliana* have resulted in somewhat well-resolved phylogenies including few accessions (BERGELSON et al. 1998; INNAN et al. 1997; VAN TREUREN et al. 1997), or trees including many accessions with minimal resolution (SHARBEL et al. 2000). Resolution in the former is likely due to the type of analysis presented; Neighbor-Joining approaches to tree building yield fully resolved trees, regardless of the level of support. Reports of low-resolution trees likely result from more rigorous analyses, but use markers with low mutation rates relative to the time scale involved.

Arabidopsis thaliana accessions are derived from natural populations that likely have histories involving interpopulation gene flow and recombination. Because of this, their reticulate evolutionary history cannot be validly represented by analyses that yield bifurcating trees. However, to provide some reference of similarity among many *A. thaliana* accessions, a cluster analysis is presented here and selections of the results are discussed below. The tree presented (FIGURE 2.4) is not proposed as a phylogeny, but

instead as a tentative framework and test of genealogical signal. This tree is a majority rule consensus of 1000 independent UPGMA runs and only shows relationships with strong support (i.e., only relationships that occur in 70% or more of all independent runs are represented), while relationships with weak support are collapsed back to a central node. The finding of strongly supported clusters and unresolved relationships between clusters likely reflects the presumed reticulate history of populations within the species and recent independent evolution of separate lineages. Below we briefly discuss a few of the more interesting results.

The relationship between the two most-utilized reference strains, Col-0 and L*er*, remains unresolved. These two accessions are purportedly derived from the same seed stock, although details of that original stock remain elusive (Nottingham *Arabidopsis* Stock Center, http://nasc.nott.ac.uk). The accumulation of mutations due to either irradiation (in the L*er* line) or generations in cultivation likely cannot explain this finding as L*er* does show strong similarity to La-0 (6765), which was also derived from the above-mentioned stock. The Col-0 genotype does not match identically with any other accessions examined in our lab. In addition, our Col-0 DNA sequences match those in the database so that seed or DNA contamination in our lab would not appear to explain this finding either. Given the low levels of both phenotypic (personal observation) and genetic similarity between Col-0 and L*er*, it would appear that the original stock was more heterogeneous than originally suspected; this is also in accord with reports of strong sequence divergence between the two accessions (e.g., Noel et al. 1999; Borevitz et al. 2003).

Not-with-standing the exception just discussed, accessions originating from a common seed stock or collection site typically cluster together (e.g., the Nok cluster). However, instances where all accessions from a locality do not cluster together (e.g., the two NW clusters) are also evident. In all, 70% of expected groupings were resolved. Again, these findings may be due to local populations that are quite heterogeneous.

Past reports on *A. thaliana* genealogies have shown little to no correspondence between geographic origin and relatedness. This has been suggested to be the result of recolonization of central and northern Europe from glacial refugia (SHARBEL et al. 2000). Although much of the consensus tree presented here shows similar incongruence between geography and genetic similarity, this finding is not ubiquitous; particular clusters show consistent biogeographic trends. For example, the "Spain" cluster illustrates close associations among many independent accessions collected from throughout Spain. Likewise, accessions from India and Tadjikistan cluster together, as do collections from several proximal geographic regions (examples are denoted with an "*" in FIGURE 2.4).

In cases where similarity is incongruent with geography, there are two likely explanations: 1) the resolved relationship is correct and explanations for the pattern observed must be sought; or 2) the genealogy is incorrect and a more appropriate marker is required. Differentiating between these two is, of course, not always simple. One approach to this problem is to seek corroborating or refuting evidence for specific genealogical hypotheses. For example, our results show strong similarity between the Br-0 (6626) accession from Czechoslovakia and Mir-0 (6798) from Italy ("Glabrous A" cluster in FIGURE 2.4). It happens that both of these lines are glabrous (lacking hairs), a relatively uncommon phenotype among wild-derived accessions. Others have reported

sequence data showing that these two accessions share the same allele at the GL1 locus (HAUSER et al. 2001), which (when knocked out) is responsible for the glabrous phenotype. In addition, there is a second microsatellite-based cluster that contains glabrous accessions ("Glabrous B" cluster), which according to HAUSER et al. (2001) share a defective GL1 locus. Taken together, there exists strong support for these particular relationships. Although not well-resolved across the entire tree, it is clear that microsatellite data possess signal useful for reconstructing the evolutionary history of this group and warrant further investigation.

Conclusions

The present analysis reveals several important aspects of microsatellite evolution and application in *A. thaliana*. Most loci examined support no one mutation model. Instead, it appears that sequence interruptions within the repeat region of microsatellite loci have a strong influence on the potential diversification of loci and should be taken into consideration in the construction of new microsatellite mutation models. Specifically, the magnitude of the effect of repeat interruptions is proportional to the lengths of the remaining intact repeat regions. Additionally, microsatellite loci of *A. thaliana* possess a high level of intraspecific phylogenetic signal. As these marker data are potentially of broad use, they can be accessed at: http://www.esb.utexas.edu/arabidopsis2010/

Chapter 3: Mapping quantitative trait loci in multiple populations of *Arabidopsis thaliana* identifies natural allelic variation for trichome density

ABSTRACT

The majority of biological traits are genetically complex. Mapping the quantitative trait loci (QTL) that determine these phenotypes is a powerful means for estimating many parameters of the genetic architecture for a trait and potentially identifying the genes responsible for natural variation. Typically, such experiments are conducted in a single mapping population and, therefore, only have the potential to reveal genomic regions that are polymorphic between the progenitors of the population. What remains unclear is how well the QTL identified in any one mapping experiment characterize the genetics that underlie natural variation in traits. Here we provide QTL mapping data for trichome density from four recombinant inbred mapping populations of *Arabidopsis thaliana*. By aligning the linkage maps for these four populations onto a common physical map, the results from each experiment were directly compared. Seven of the nine QTL identified are population-specific while two were mapped in all four populations. Our results show that many lineage-specific alleles that either increase or decrease trichome density persist in natural populations and that most of this genetic variation is additive. More generally, these findings suggest that the use of multiple populations holds great promise for better understanding the genetic architecture of natural variation.

INTRODUCTION

The genetic and molecular bases of complex traits are poorly understood. A common approach to this problem is the use of whole-genome scans to identify polygenes, or quantitative trait loci (QTL) (ABIOLA et al. 2003; BARTON and KEIGHTLEY 2002). The results of such analyses provide estimates of several genetic parameters that underlie phenotypic variation, including the number of loci, the type and magnitude of their effects, interactions between genes (epistasis), and gene-by-environment interactions – collectively referred to as the genetic architecture of a trait. However, these parameters are largely population-specific (LYNCH and WALSH 1998) and most species show some degree of population genetic structure (Hamrick and GODT 1996; Bohonak 1999). Therefore, any description of a trait's genetic architecture made from a single population, natural or experimental, likely describes only a small part of what might be thought of as the "global genetic architecture", or all of the loci, their effects, and potential interactions that contribute to standing variation for a trait within a species. Although certain to vary among traits and organisms, in general, we have very little information regarding how well any one mapping study captures this global genetic architecture.

Our best initial opportunities to address this issue lie with highly variable traits in model organisms, where strong genetic infrastructures have been established (e.g., see FRARY et al. 2000; EL-DIN EL-ASSAL et al. 2001; ROBIN et al. 2002). To this end, trichome density in *Arabidopsis thaliana* is one such trait (Fig. 3.1, data collected in the present study). Trichomes (plant hairs) are present on nearly all land plants and are

known to play important roles in plant protection, specifically against insect herbivory (POOS 1929; GILBERT 1971; KENNEDY 2003), drought (SMITH and NOBEL 1977; SANDQUIST and EHLERINGER 2003), and UV radiation (KARABOURNIOTIS and MANETAS 1995; LIAKOURA et al. 1997). The near ubiquity of trichomes and their morphological diversity argue for a strong role for selection in their maintenance and ultimate diversification (see LEVIN 1973 for review).

In *A. thaliana*, the single-celled trichome has become a popular system for the study of cellular differentiation and development processes (SZYMANSKI et al. 2000; LARKIN et al. 2003). Many of the genes involved in trichome initiation and development have been identified through forward genetics (KOORNNEEF 1981; KOORNNEEF et al. 1982; HULSKAMP et al.1994; ZHANG et al. 2003) and have led to hypothetical genetic models for the pathways that control trichome development (LARKIN et al. 2003; ZHANG et al. 2003). Despite this body of work, little is known about the genetics that underlie standing variation for trichome density, a character known to be of selective importance in natural populations of *A. thaliana* (MAURICIO 1998). The only published study of the quantitative genetics of trichome density in *A. thaliana* (LARKIN et al. 1996) provided mapping results from one recombinant inbred line (RIL) population and identified a single QTL of very large effect (73% of within population variation). It is highly unlikely, however, that a single gene is responsible for the twenty-fold range of natural variation observable in trichome density within this species (Fig. 3.1). To more thoroughly explore the genetic basis of trichome density variation and the potential of multiple-population mapping, we have mapped QTL for this trait in four RIL populations derived from six *A. thaliana* accessions. Our mapping results identify nine QTL, termed

Trichome Density Loci 1-9. Here we provide mapping results, comparative analyses, and interpretations of our findings.

MATERIALS AND METHODS

Natural variation survey for trichome density

To screen natural variation for trichome density within *A. thaliana*, 79 single seed descent accessions that represent populations from throughout the species' range were acquired from the *Arabidopsis* Biological Resource Center (ABRC). Three replicates of each accession were planted in 6.35 cm pots (one plant per pot) in Promix BX soil (Hummert International), randomized across flats, and vernalized for seven days at 4°. Flats were then moved to 20° under 24 hour light. At 21 days post-germination, the fully expanded third true leaf of each plant was removed and all trichomes on the adaxial surface were counted under a dissecting microscope at 35X magnification. Each leaf was then pressed flat for 24-48 hours, digitally photographed, and the total leaf blade area measured using ImageJ v1.32 (NIH Image). The number of trichomes per square centimeter was calculated for each leaf, and the mean density for each line was determined. Total phenotypic variance was partitioned into genotypic and environmental (error) components by ANOVA.

Mapping populations

Four recombinant inbred line (RIL) mapping populations were used in this study. Seed for three populations, L*er*-0 x Col-4 (LISTER and DEAN 1993), L*er*-2 x CVI

(ALONSO-BLANCO et al. 1998), and Bay x Sha (LOUDET et al. 2002), were acquired from the ABRC. The fourth RIL population, L*er*-2 x No-0, was developed jointly by the Lloyd (UT-Austin) and Casal (Universidad de Argentina) labs. This population was genotyped using microsatellite loci and both genotyping data and seed will be made publicly available in the near future. All four populations were genotyped in the F8 generation or later, yielding lines that are greater than 99% homozygous; genotyping details for each population can be found in the original descriptions cited above and at The *Arabidopsis* Information Service (http://www.arabidopsis.org) and Nottingham *Arabidopsis* Stock Center (http://nasc.nott.ac.uk) web sites. The numbers of RILs and markers used for mapping in each population are listed in Table 3.1. This set of four populations has the additional benefit of a test-cross design. Because three of the populations share the L*er* accession as a parent, their results can be directly compared in reference to a tester line.

Scoring the trichome density phenotype

For each mapping population, five replicates of each RIL were randomized across PRO72 multicell flats filled with Promix BX soil, vernalized for 11 days at 4°, and ultimately maintained under 14:10 (light:dark) days in a glasshouse at Brackenridge Field Station in Austin, TX until time of data collection. All experiments were completed between March and May of 2003; although growing each population concurrently is ideal, the scale of each mapping experiment precluded this. At approximately 24 days post-germination, the fifth true leaf from each plant was removed and the number of

trichomes was counted within a 25 mm^2 area midway between the midrib and leaf edge at the widest point on the adaxial leaf surface.

Linkage map construction

Because of differences in marker use between this study and others, all linkage maps were regenerated from raw marker data. Markers for each population were selected from all available data sets to optimize genome coverage and reduce missing genotypes. Linkage maps were constructed using JoinMap 3.0 (VAN OOIJEN and VOORRIPS 2001) and were considered complete only when unambiguous marker order was attained.

QTL mapping

The mean trichome density score was calculated from replicates of each RIL and used for mapping. QTL were mapped for each population independently. To identify regions of the genome that possess significant QTL, we used interval mapping and Multiple QTL Model (MQM) mapping (akin to composite interval mapping) as implemented by the MapQTL 4.0 software package (VAN OOIJEN et al. 2002). QTL significance thresholds were determined empirically for each mapping population by permuting the trait data over individuals while holding the marker data fixed. One thousand iterations provided a null distribution, from which a 95% significance threshold was attained. Mapping was conducted via an iterative method. Interval mapping was performed first to identify QTL with significant main effects. The markers most closely associated with each QTL were then selected as cofactors for MQM mapping. If new QTL were identified under the new model, their associated markers were incorporated as

cofactors for the next round of mapping until no new QTL were identified. MQM mapping performs the same way as interval mapping while statistically controlling for the effect(s) of other QTL (cofactors). The only exception to the MapQTL manual-suggested method of mapping is that when estimating the LOD score and confidence intervals for any one QTL, markers flanking the QTL of interest were deselected as cofactors. The effect of each QTL was estimated in MapQTL and reflects the percentage of variance within a population that is explained by the QTL. Hereafter, when QTL "effects" are discussed, they refer to this measure.

A screen for QTL that interact epistatically was performed using the Pseudomarker multiple-QTL framework presented by Sen and Churchill (SEN and CHURCHILL 2001). This methodology employs a Monte Carlo imputation algorithm to simulate multiple versions of complete genotype information on a dense genome-wide grid; these imputed genotypes are referred to as "pseudomarkers". The pseudomarker grid was scanned using both one and two QTL models at each position across the genome and evidence for a QTL or for QTL-QTL interactions (epistasis) was determined using robust 1- and 2-dimentional permutation tests. Details of model selection and QTL mapping for epistasis detection followed that of JUENGER et al. (2004).

Map alignments

Differences in linkage map length between populations are generally expected due to nonoverlapping markers, different marker densities, and presumed variation in recombination rates (SANCHEZ-MORAN et al. 2002). As a consequence, comparing mapping results from different populations is difficult. To align the four linkage maps

per chromosome that were generated here, the linkage positions of markers in each population were aligned with their corresponding physical positions on the Col physical map (ARABIDOPSIS GENOME INITIATIVE 2000). To more closely compare QTL positions across mapping populations, we also estimated the physical positions of QTL based on the coarse linear relationship between linkage map positions and physical positions for each marker using a least squares regression approach. Although there is certain to be intermarker variation for recombination rates (COPENHAVER et al. 2002), QTL physical position estimates are bound by known flanking marker alignment positions. For ease of interpretation, QTL mapped in the L*er*-0 x Col-4 population were plotted directly on the physical map. QTL mapped in different populations were inferred to be the same locus when the physical positions of markers within 2-LOD support intervals overlapped (see results for assumption details).

RESULTS

Natural variation for trichome density

To examine the distribution and nature of trichome density variation within *A. thaliana*, 79 natural accessions were screened. ANOVA revealed highly significant among-line variance ($p \leq 0.0005$), indicating that a strong genetic component underlies variation for trichome density within *A. thaliana*. The broad sense heritability estimate (V_G/V_P) for trichome density in this collection of lines was 0.703. Excluding completely glabrous lines, the results of this screen showed there to be a roughly 20-fold range of variation for trichome density that has a broad, normal to bimodal distribution (Fig. 3.1).

Genetic analysis of mapping populations

ANOVA showed there to be significant variation for trichome density among RILs within each of the four mapping populations ($p \leq 0.0001$ for each population). Consistent with this, heritability estimates (V_G/V_P) were relatively high in all mapping populations, ranging from 0.795-0.896 (Table 3.1). These heritabilities are also consistent with that estimated from natural accessions. The frequency distributions for trichome density were relatively normal for each mapping population (Fig. 3.2); the three populations that have L*er* as a parent show roughly the same distribution, while the Bay x Sha population had a similar spread, but the entire distribution was shifted higher by ~10 trichomes.

Linkage maps and alignments

Separate linkage maps were constructed for each mapping population. The use of different markers, numbers of markers, and presumed differences in recombination rates between populations led to considerable variation in total linkage map lengths, which ranged from 364-490 cM (Table 3.1). Because of this variation, all linkage maps for a given chromosome were aligned using the physical positions of markers for each population (see below).

Mapping results

Three to five QTL with significant effects on trichome density were identified in each mapping population. Initially, a total of 15 QTL were mapped, with individual effects that explained from 6.8-25% of within-population variation (Table 3.2). Models

including gene-by-gene interaction showed the presence of only weak epistasis (4% effect) between two loci in the L*er*-2 x CVI population (Figure 3.3). Although the two parents for the Bay x Sha population failed to germinate, thus preventing an analysis of transgressive segregation, transgressive phenotypes were observed in the other three populations. Strong genetic evidence for transgressive segregation (based on allelic effects at mapped QTL) was only clear in the L*er*-2 x CVI and L*er*-2 x No-0 populations (Table 3.2), where both parent accessions possess high and low alleles for trichome density. In the L*er*-0 x Col-4 and Bay x Sha populations, QTL effects were unidirectional; one parent possesses the low allele at all loci (L*er*-0 and Bay) and the other parent has all high alleles (Col-4 and Sha).

QTL overlap between populations

To investigate whether the same or different QTL were mapped in the four RIL populations, linkage map positions were aligned with known physical positions for markers and with estimated physical positions for QTL (Fig. 3.4). In general, there is strong overlap for QTL positions and 2-LOD support intervals at two loci across all four populations (bottom of chromosome two and top of chromosome four). Although in a different context (mapping QTL for different traits in the same population), the occurrence of overlapping support intervals is often interpreted as evidence that either the same QTL was mapped or that genes contributing to trait variance are positionally clustered. Because of the observed overlap and unidirectional allelic effects (L*er* has the "low" allele at both loci that show overlap across populations), it seems likely that the same QTL was mapped in all four populations (considered further in discussion). The

remaining seven QTL demonstrate no positional overlap among populations. Thus, the initial 15 QTL were condensed to nine independent QTL, named TDL1-9.

DISCUSSION

Although the mapping of quantitative trait loci in experimental populations has become a common method by which many parameters of a trait's genetic architecture are characterized (ABIOLA et al. 2003), the extent to which these specific parameters differ across populations remains largely unexplored. Here, we have broadly surveyed natural lineages of *A. thaliana* for trichome density and detected a roughly 20-fold range of variation. To better characterize the genetics that underlie this variation and to investigate genetic architecture heterogeneity, we mapped QTL in four recombinant inbred line populations. The results from each population were compared by aligning the linkage maps for each chromosome across all four populations. Our results identify at least nine QTL with significant effects on trichome density; two QTL were mapped in all four populations and seven were mapped in one population each. These data allow us to 1) characterize the nature of quantitative genetic variation for trichome density in *A. thaliana*, 2) compare the results of mapping QTL in multiple populations derived from different individuals, and 3) initiate a search for candidate genes that may play a role in standing variation for the trait.

Trichome density QTL study comparison

A total of nine QTL that affect trichome density were identified here, with at least one QTL occurring on each of the five *A. thaliana* chromosomes (Table 3.2 and Fig. 3.4).

These results are in sharp contrast to those reported by LARKIN et al. (1996) in the only related effort to map QTL for trichome density in *A. thaliana*. Their study focused on the number of trichomes that develop on first true leaves in a single RIL population, L*er*-0 x Col-4, and identified one QTL (Reduced Trichome Number, RTN). The reported map position of RTN overlaps with a QTL for trichome density identified here, TDL2. This locus was mapped in all four populations examined, including the same L*er*-0 x Col-4 population studied by LARKIN et al. (1996), but the effect estimated in this study (11.5-21% of within population variation explained) is much smaller than that estimated for RTN (73%). In addition to the TDL2/RTN locus, our mapping results from the L*er*-0 x Col-4 population identify two other QTL for trichome density (Table 3.2 and Fig. 3.4). The differences between the two studies in the number of QTL mapped and their estimated effects may be due to differences in the phenotype scored (number of trichomes on first true leaves (LARKIN et al.) versus trichome density on fifth true leaves (this study)), methods of analysis, different environmental conditions or some combination of these. Although any QTL mapping study can only identify a minimum number of loci that affect a trait, our findings highlight the power afforded by mapping in multiple populations. Our results from the other three populations identify six more QTL, for a total of eight not previously described for trichome density, several of which are of relatively large effect.

Overlapping QTL

The map positions and 2-LOD support intervals for QTL in only two genomic regions (bottom of chromosome two (TDL2) and top of chromosome four (TDL5)) show

strong overlap across all four populations studied (Fig. 3.4). Many scenarios may explain this result; at each locus, there may be as many as six or as few as two alleles segregating among the parental accessions. The most parsimonious explanation for overlapping QTL across these populations is that the L*er* accession possesses an allele that differs in effect significantly from the Col-4, CVI, and No-0 alleles, all of which may or may not possess the same allele at each locus. This scenario is supported by the fact that the L*er* allele confers lower trichome density at TDL2 and TDL5 in all three populations for which it is a parent. Whether the same or different alleles are segregating in the Bay x Sha population as those in the other three populations is not clear. However, the effect of the TDL5 locus in the Bay x Sha population is more than double that of any of the other three (Table 3.2), suggesting that, if it is the same locus, then the allele(s) present are different or the effects of the alleles are conditional on genetic background.

Population-specific QTL

The genetic architecture elucidated for any trait is expected to be population-specific. While among-population variation for genetic architecture has been detected by classic quantitative genetic approaches (e.g., DONOHUE et al. 2000; MORGAN et al. 2001; ASHMAN 2003), little has been described on exactly *how* genetic architecture for a trait varies among natural or experimental populations; specifically, we would like to know which components of genetic architecture vary and in what way(s). Although the experimental design employed here precludes an examination of every aspect of genetic architecture, our analyses reveal the strongest evidence of genetic architecture variation by the mapping of different QTL in different populations (Table 3.2 and Fig. 3.4).

Differences in the QTL identified for a trait among RIL populations may be due to differences in the power to detect QTL, environmental variation between mapping experiments, or the segregation of meaningful genetic variation among natural lineages. Because population-specific QTL of relatively small effects were mapped in all four populations, it does not appear that RIL population size or marker coverage contributed significantly to the mapping of different QTL in different populations. To investigate the potential for a significant gene-by-environment component, subsamples of each population were grown concurrently several months after the initial experiments and scored in the same manner as before (data not shown). Comparisons between initial and subsequent RIL means showed no significant differences (paired t-test, p=0.355). Taken together, these analyses suggest that the population-specific QTL detected are due to natural genetic variation for trichome density that is partitioned among accessions of *A. thaliana*. This is further supported by the broad distribution of natural trichome density phenotypes and a high among-accession heritability estimate (Figure 3.1).

What do these mapping population-specific QTL mean? QTL are identified based on a significant difference in the mean phenotype of RILs possessing alternative alleles. Therefore, any QTL identified in only one of the four mapping populations is likely due to an accession-specific allele, otherwise the polymorphism also would have been mapped in other populations. This hypothesis and the origin of novel alleles are most easily explored by, again, comparing mapping results from the three populations that share the L*er* accession as a parent; these three populations are, effectively, a set of test-crosses. For example, TDL3 was mapped only in the L*er*-0 x Col-4 population (Fig. 3.4). Because this locus was not mapped in the other two populations, the L*er* allele at

this locus is functionally, perhaps also practically, the same allele as those possessed by CVI and No-0. Therefore, the Col-4 accession must possess an allele at the TDL3 locus that is not present in either of the other mapping populations. By this reasoning, any QTL mapped in one of these populations and not the other two represents an allele sampled only in the non-L*er* accession of the population in which the QTL was mapped. Because the Bay x Sha population does not share a parent with any other mapping population, results from this population cannot be compared in the same way as above. However, the TDL4 locus was mapped only in this population, so at least one of these two accessions must possess an allele not sampled in the other mapping populations. In all, seven population-specific QTL were mapped, each of which reveal an accession-specific allele that either increases (TDL1, 3, and 9) or decreases (TDL6, 7, and 8) the trichome density phenotype relative to L*er*; the effect of the novel allele(s) at TDL4 cannot be inferred from the data.

Epistasis

Among the QTL identified here, significant gene-by-gene interaction (deviation from an additive model) was detected between only two loci in only one population (L*er*-2 x CVI); this interaction is responsible for four percent of within population variation, which is less than that of any main effect QTL. Interestingly, one of these QTL was mapped in all four populations (TDL2) and the other was identified only in the L*er*-2 x CVI population (TDL6). In general, it seems that the majority of genetic variation for trichome density within *A. thaliana* is additive in nature, however, it should be noted that

the Pseudomarker analyses performed are very conservative and may underestimate the epistatic contribution to trichome density variation.

Transgressive segregation

Recently it was argued that the genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations (RIESEBERG et al. 2003). Our data show that in the three populations for which parent phenotypes were observed, transgressive segregation was evident for trichome density (Fig. 3.2). However, only two of the four mapping populations showed strong genetic evidence of transgressive segregation (L*er*-2 x CVI and L*er*-0 x No-0); that is, for only two populations does each parent possesses at least one "high" and one "low" trichome density allele among the detected QTL. The observation of transgressive phenotypes in populations where QTL allelic effects are unidirectional may reflect QTL of relatively small effects that were undetected.

Although all populations studied here are experimental, they were derived from (mostly) natural lineages of *A. thaliana* and these results provide some insight regarding the potential for transgressive segregation in natural populations. Specifically, relative to one another, at least three of the six parent accessions possess a genetically unique way of increasing trichome density and at least three possess a genetically unique way of decreasing trichome density. So, even though all four experimental populations do not show strong genetic evidence of transgressive segregation, the finding of several lineage-specific alleles that alter trichome density in the same way shows that sufficient

intraspecific genetic variation exists to generate this phenomenon in natural populations of *A. thaliana*.

Candidate loci and phenotype regulation

The complete genome sequence of *A. thaliana* and a rich trichome development literature enabled the search for candidate genes near mapped QTL. For four of the nine QTL mapped here, strong candidate genes lie within or very near ascribed confidence intervals (Table 3.2, Fig. 3.4). One of the QTL maps near a gene for which knock-outs have known quantitative effects on trichome density (TDL9), and three map near known regulators of trichome initiation (TDL2, 3, and 8). Notably, all of the *A. thaliana* genes identified by conventional genetics that affect trichome initiation, including these four candidate loci, either encode proteins with transcription factor motifs or a protein that directly interacts with transcription factors (LARKIN et al. 2003; ZHANG et al. 2003). Although striking, this finding is in accord with early and recent speculation that quantitative trait variation results from regulatory variants more often than from coding sequence variants (KING and WILSON 1975; MACKAY 2001; KORSTANJE and PAIGEN 2002); these suppositions focus mainly on the contrast between mutations within regulatory regions versus mutations within coding regions. However, recent evidence shows that protein-coding change in transcription factors may play a larger role in pathway evolution than previously thought (HSIA and MCGINNIS 2003). Therefore, this contrast may be quite naturally extended to include mutations in regulatory regions *and* mutations within coding regions of regulatory genes versus coding mutations in structural genes alone. Indeed, of the ten QTL thus far cloned from plants (PARAN and ZAMIR

2003), at least one-half are either transcription factors or the phenotype has been ascribed to altered gene expression levels. Of course, precisely determining the genetic sources of variation in trichome density will require the identification and molecular characterization of mapped QTL, and each of the candidate loci identified here provide an opportunity to do just this.

Conclusions

There are many hypotheses regarding the genetic architecture of natural trait variation, and it seems clear that the correct hypothesis will depend on the particular trait and organism under study. Very generally, standing variation for quantitative traits may be due to few genes with many mutations - each with their own distribution of effects, several genes with few or many segregating mutations with intermediate to small effects, or some combination of these. As these parameters apply directly to the potential for selection response and morphological evolution in general (BARTON and KEIGHTLEY 2002), understanding the frequencies at which different genetic architecture hypotheses hold true is an important goal. Among the most common methods for estimating parameters of genetic architecture is QTL mapping. However, the majority of QTL studies in natural systems utilize a single population derived from only two individuals (see LONG et al. 1995 and GURGANUS et al. 1999 for exceptions) and an obvious, though typically understated, caveat to these studies is that they characterize only some portion of the genetic architecture that underlies trait variation. Importantly, we do not yet know how well the results from any one mapping population describe the genetics that underlie standing variation within a species. If we intend to use QTL mapping to understand the

range of genetic variation that is responsible for standing trait variation, a broader approach is needed.

To explore the extent of interpopulation variation for genetic architecture, we have mapped QTL for trichome density in four RIL populations of *A. thaliana*. Where only a single QTL had been described for this trait previously, our results identify a total of nine QTL. Comparative analyses reveal that much of the genetic variation detected for trichome density is partitioned among lineages of *A. thaliana*. Specifically, these data reveal multiple accession-specific alleles that confer similar directional change; whether or not this pattern extends to a broader sampling of *A. thaliana* lineages will be an interesting line of inquiry. More generally, our findings show that QTL mapping results can be strongly dependent upon the particular population(s) in which one maps, and that a multiple mapping population approach has the potential to much better characterize the species-wide genetic architecture for a trait. Fortunately, for several model systems, the development of new mapping populations is under way, which will provide opportunities to more broadly characterize the genetic architecture of complex traits and eventually aid in identifying the molecular variants responsible for natural variation.

Chapter 4: *Atmyc1*: a QTL for trichome density in *Arabidopsis thaliana*

INTRODUCTION

Most biological traits exhibit a continuous or quantitative distribution due to polygenic control. Characterizing the genetic architecture of quantitative traits remains an important step toward understanding the molecular bases of biological diversity and the evolutionary forces under which it is generated and maintained. A common first approach toward this goal is the identification of quantitative trait loci (QTL) by association mapping (BARTON and KEIGHTLEY 2002). Such analyses provide minimum estimates of the number of loci affecting variation in a trait, the relative magnitudes of their effects, and, depending on the experimental design employed, the relative contributions of epistasis, pleiotropy, dominance, and gene-by-environment interactions (FALCONER and MACKAY 1996). Over the past 15 years, advances in molecular marker technologies have greatly increased the number of organisms in which QTL mapping can be accomplished and broadened the types of questions that can be addressed using this approach (reviewed by (ERICKSON et al. 2004)). Beyond mapping QTL, the next step is to identify the specific molecular variants that fuel phenotypic variation (ABIOLA et al. 2003; PARAN and ZAMIR 2003). Progress here has been slow and success remains limited to relatively few traits in only a handful of organisms (GLAZIER et al. 2002). In part, the limited number of cloned QTL can be attributed to a lack of genomic data for most organisms; however, even in model systems, identifying the gene(s) that underlie mapped QTL has proven a difficult and slow process.

To make the difficult leap from QTL to gene and, ultimately, to the specific polymorphism(s) that underlie trait variation, two general approaches have been utilized: fine mapping and the candidate gene approach. Fine mapping is the process by which QTL intervals are narrowed by successively recombining away the genomic regions surrounding the QTL (e.g., constructing near isogenic lines), until a small enough interval remains to allow a switch to molecular approaches, such as chromosome walking. The candidate gene approach can either follow fine mapping or be initiated straight away (HORIKAWA et al. 2000), but requires vast genomic resources - genomic regions of interest (that underlie QTL intervals) can be scanned and candidate genes nominated for in-depth molecular analysis. Clearly the candidate approach is currently limited to model organisms, and will have the highest likelihood of success for highly variable traits, for which traditional genetic approaches have identified potential candidate genes.

Trichome density in the model flowering plant, *Arabidopsis thaliana*, demonstrates a very broad distribution and is known to be under strong genetic control (H^2 ranges from ~0.73 – 0.89). Previous QTL mapping results for trichome number (LARKIN et al. 1996) and trichome density (SYMONDS et al. in press) identified one and nine QTL, respectively. The discrepancy in the number of QTL identified is likely due to the scoring of different trichome phenotypes in different mapping populations; Larkin et al. used the L*er* x Col population and Symonds et al. used the L*er* x Col, L*er* x CVI, L*er* x No-0, and Bay x Sha populations. Two of the nine QTL described by Symonds et al., were mapped in multiple populations. One of these, TDL5, was mapped to the top of chromosome four independently in each of the four populations (FIGURE 4.1). Estimates of the physical position of this QTL and the two-LOD support intervals that surround

them overlapped completely among populations, suggesting that the same locus was mapped independently in each population. In an initial screen of the region, no candidate gene with a known mutant trichome phenotype was discovered, however, the search did reveal a bHLH-containing gene, *ATMYC1*, two paralogs of which (HEIM et al. 2003; TOLEDO-ORTIZ et al. 2003) are known to have trichome phenotypes upon knockout (PAYNE et al. 2000; ZHANG et al. 2003). *ATMYC1* has been described as being expressed in seeds (URAO et al. 1996), but no trichome phenotype has been reported. Here, we present results from complementation and molecular variation approaches that indicate that *ATMYC1* is a QTL for trichome density.

MATERIALS AND METHODS

The *ATMYC1* mutant (*atmyc1*) phenotype

A TDNA insertion line (SALK_0057388) for the *ATMYC1* locus (At4g00480) was obtained from The *Arabidopsis* Information Resource (ABRC). This line was generated in the Col-0 background (http://signal.salk.edu/cgi-bin/tdnaexpress) and is homozygous for a TDNA insertion that resides in the 5' flanking region of the gene, that presumably disrupts transcriptional regulation. Relative to the Col-0 background accession (~30 trichomes per first true leaf), the *atmyc1* mutant line was determined to have a strongly reduced trichome density phenotype; ~16 trichomes per first true leaf versus ~30 on wildtype Col-0. This mutant line was used in subsequent complementation tests (described below).

Quantitative complementation tests

To test the hypothesis that the *ATMYC1* locus is responsible for trichome density variation mapped in previous QTL studies (SYMONDS et al. in press), quantitative complementation tests were performed. Because the *atmyc1* mutant is in the Col-0 background, the most direct comparison that could be made (with regard to QTL mapping results) was between Col-0 and L*er*-0; the L*er* x Col mapping population is the only one in which the Col accession is a parent. Unless stated otherwise, references to the Col and L*er* accessions are Col-0 and L*er*-0, respectively. Pollen from the L*er* accession was crossed onto the stigmatic surface of emasculated flowers of the *atmyc1* mutant and of the Col wildtype. To control for potential cytoplasmic variation among accessions, all crosses were made in the direction of Col or *atmyc1*. The resulting F_1s differ only at the *ATMYC1* locus. This allowed for comparisons between individuals with a Col/Ler and an *atmyc1*/L*er ATMYC1* genotype, while holding the rest of the genome constant. That is, the only difference between the two sets of progeny is the replacement of the Col allele with a null (mutant) allele. To test for a dosage effect, the Col individual was crossed to *atmyc1*, which yields a Col individual with a single functional *ATMYC1* allele (*atmyc1*/Col). To explore allelic effects of the *ATMYC1* locus in a broader sampling of accessions, the same crossing scheme described above was concurrently carried out for the remaining four parent accessions used in the previous mapping study (No-0, Bay, CVI, and Sha). Fifteen replicates of each F_1 genotype, parental accession, and the *atmyc1* mutant were potted in five pots and the pots were randomized across three flats. All individuals were vernalized in the dark for four days at 4°C, and subsequently moved to a fluorescently lit 20°C growth chamber. Upon emergence of the

fifth and sixth true leaves the number of trichomes on each of the first two true leaves of each seedling were counted under 50X magnification on a dissecting microscope. The mean for each individual was determined and each genotype mean calculated from these. In previous work, we scored a trichome density phenotype on fifth true leaves. Because trichome density on first true leaves is correlated with trichome density on later leaves, and because divergent phenotypes are more obvious on first leaves, we switched to scoring the number of hairs on first true leaves for this work.

Transgenic complementation tests

To test for functional differences between natural alleles of *ATMYC1*, genomic alleles with native promoter regions from four accessions were transformed into the *atmyc1* mutant background. Primers were designed from the Col-4 *ATMYC1* sequence to anneal 1.7 kb upstream and 1.6 kb downstream of the start and stop codons, respectively. Each primer had either the BP-1 or BP-2 tail required for use in the Gateway recombination cloning system (Invitrogen, Inc.). These primers were used to PCR amplify and clone the genomic allele with native promoter elements from each of the Col, L*er*, Sha, and CVI accessions used in the four mapping populations described previously. These four accessions were chosen based on sequence divergence; Col and CVI have very similar alleles, which differ substantially from the L*er* and Sha alleles (see section on molecular evolution of *ATMYC1*). Each PCR product was cloned into the pDONR-Zeo plasmid using a recombinase system (Invitrogen) and transformed into DH10B *E. coli* cells. Clones were screened initially by PCR, followed by sequencing. One of each of the correct *ATMYC1* alleles was then recombined into the pBAR-rfb

TDNA vector and again recovered in DH10B cells. Each of the correct constructs were subsequently transformed into agrobacterium strain GV3101, which facilitated vacuum infiltration transformation into *atmyc1* mutant plants.

Fully developed seed were harvested from each of the transformed plants and allowed to further dry down for 10 days. Approximately 500 seeds from each transformed pool of plants were potted in each of three 6.35 cm square pots with Promix BX soil. Because the vector TDNA contains a herbicide (BASTA) resistance gene along with the target DNA, selection for positive transformants was conducted post-emergence in soil. At five days post-emergence (cotyledon stage), 20mL (0.9 μg/mL) of BASTA was top-watered into each pot. Because BASTA is both a contact and systemic herbicide, selection is quite strong and only positive transformants survive. The effects of the herbicide treatment became clear at ~4 days post-treatment, whereupon the nonresistant (untransformed) seedlings became chlorotic and gradually dried completely, whereas the resistant (transformed) seedlings continued to develop normally.

Once positive transformant seedlings developed fifth and sixth true leaves, the number of hairs on each of the first true leaves was counted for each plant under 50X magnification. The average number of trichomes for the first two leaves of each plant was determined and recorded. Because within-construct distributions were non-normal, significant differences among transformation constructs were tested by a Kruskall-Wallis test, and significantly different pairs of transformant pools identified using the non-parametric Tukey's HSD post-hoc test (SOKAL and ROHLF 1997).

Natural survey of allelic variation for *ATMYC1*

DNAs were isolated from 73 natural accessions and three lab strains of *A. thaliana* (ABRC), as well as from four outgroup taxa (*Arabidopsis lyrata* (provided by Neil Harriman), *Arabis griffithiana* (ABRC), *Crucihimalaya himalaica* (ABRC), and *Capsella bursa-pastoris* (collected on UT-Austin campus)) following a modified CTAB method (Doyle and Doyle 1987). Primers were designed from the Col-4 *ATMYC1* sequence (Genbank accession #NC003075) to PCR amplify the open reading frame plus ~200 bp up- and down-stream of the start and stop codons. These primers were used with manufacturer-supplied 1X Taq buffer, 1 U AccuPrime High-Fidelity Taq polymerase (Invitrogen), and ~20 ng genomic DNA. PCR amplification was carried out in 96-well plates. Samples were checked for amplification success on 0.7% agarose gels stained with ethidium bromide, and were subsequently purified in Multiscreen PCR cleanup plates (#MANU03050, Millipore). Because flanking primers would not amplify the *ATMYC1* locus from the outgroup taxa (presumably due to sequence variation within priming sites), primers that anneal to the first and last 21 bp of the gene were used for these four species. Approximately 100 ng of each purified PCR product were then used in each of seven sequencing reactions using primers designed to anneal at staggered internal positions, providing a minimum of two overlapping sequences across the entire gene. Allelic contigs were constructed and sequence editing and validation were performed using Sequencher v.4.2.2 (Gene Codes Corp.). Full length genomic sequences for all accessions and outgroups were aligned initially using ClustalX v.1.83 (THOMPSON et al. 1997), and subsequently corrected by hand. To generate cDNA sequence alignments, flanking sequence and introns were identified using the published Col-4

cDNA sequence as template (Arabidopsis Genome Initiative). DNA flanking the open reading frame and introns were removed from the alignment and converted to cDNA sequences using BioEdit v.5.0.9 (HALL 1999).

Distance-based cluster analyses were performed on the aligned cDNA sequences using PAUP* (SWOFFORD 2003). A neighbor-joining tree was generated using Jukes-Cantor distances. Branch support was assessed by 1000 bootstrap replicates with complete character replacement using the same Neighbor-joining options used to generate the tree. The resulting gene tree was rooted using the four outgroups, *Arabidopsis lyrata*, *Arabis griffithiana*, *Crucihimalaya himalaica*, and *Capsella bursa-pastoris*.

Molecular evolution of the *ATMYC1* locus

To examine nucleotide diversity and molecular evolution of the *ATMYC1* locus, the sequence analysis software DNASP v.4.00 (ROZAS et al. 2003) was used. The common nucleotide diversity indices, π (NEI 1987) and θ_w (WATTERSON 1975), were measured across the entire gene. To assess intra-gene variation for nucleotide diversity, a sliding window analysis was run along the full length (start to stop) genomic sequence alignment of the 76 *A. thaliana* alleles. Window length was set at 50 bp and moved along the alignment at 10 bp intervals. Because of initial observations of high levels of diversity and divergence among *ATMYC1* alleles, we tested the null hypothesis of neutral molecular evolution at this locus by comparing the number of nonsynonymous substitutions per nonsynonymous site (K_A) to the number of synonymous substitutions per synonymous site (K_S). Synonymous and nonsynonymous mutations are thought to

arise at roughly the same rate, but nonsynonymous changes are more likely to affect protein function. Therefore, nonsynonymous substitutions with positive (selectively beneficial) effects are more likely to be swept to fixation than neutral mutations. By examining the ratio of K_A/K_S, one may identify signal indicative of this (KIMURA 1977). K_A/K_S ratios near one are thought to be indicative of a neutrally evolving gene or region of a gene, values significantly below one are purported to be under purifying selection, and values significantly greater than one indicate positive selection. Because different regions of a gene may have experienced different forms of selection, a sliding window analysis was used to examine sequence divergence of *ATMYC1* alleles; the window size was set at 99 bp, and was slid in 15 bp increments along the length of aligned cDNA sequences. These ratio plots were generated in two ways: (1) using local K_A over local K_S measures (traditional method) and (2) using local K_A values over the gene-wide average K_S value. While the former method is the convention, the latter has been suggested as an alternative to deal with false or misleading positives brought on by very low local K_S values (SHIU et al. 2004). For each window of sequence the K_A/K_S ratio was calculated and plotted. These comparisons focused on divergence between Type I and Type II alleles (alleles were assigned to Type I or Type II based on gene tree results below).

RESULTS

Quantitative complementation tests of *ATMYC1*

The quantitative complementation tests were revealing in several ways. First, the test cross of Col x *atmyc1* showed no evidence of a gene dose effect (FIGURE 4.2). That is, the Col x *atmyc1* trichome phenotype does not differ significantly from that of the Col wildtype. Therefore a single functional allele of Atmcy1 is sufficient to recover the mutant phenotype. In reference to previous mapping results and whether or not *ATMYC1* is one of the QTL identified, the most direct comparison to be made is between the Col and L*er* alleles. Similar to the Col x *atmyc1* phenotype, Col x L*er* also has a phenotype nearly identical to that of the Col wildtype. Since we know that Col possesses all high alleles for trichome density relative to L*er* (SYMONDS et al. in press), this result indicates that Col alleles for trichome initiation are dominant to L*er* alleles. In contrast, the lower trichome phenotype of *atmyc1* x L*er* shows that a single copy of the L*er* allele is insufficient to recover the *atmyc1* mutant phenotype. In fact, *atmyc*1 x L*er* has a phenotype very similar to *atmyc1*, placing L*er* in the same complementation group as *atmyc1*. This is strong evidence that variation at *ATMYC1* is responsible for some portion of the trichome density variation segregating in the Col x L*er* mapping population.

For the remainder of the crosses, interpretations are difficult for two reasons. Unlike for Col x L*er*, in all other pairwise crosses, we do not know: (1) what or how many QTL for trichome density may be segregating or (2) the patterns and potential roles of dominance and epistasis. Generally, however, all F_1s made with *atmyc1* demonstrate

lower trichome phenotypes than corresponding F_1s made with Col, perhaps suggesting that the Col allele may be much better at its role in the trichome initiation pathway than all other alleles tested. Alternatively, a gene dose effect may be at play, but the magnitude of its effect is dependent upon genetic background.

Transgenic complementation tests

Genomic *ATMYC1* alleles from the Col, L*er*, CVI, and Sha accessions were transformed into the *atmyc1* mutant background. From twelve to sixteen independent transformants were recovered for each allele and, as is often the case, a range of phenotypes were observed among the independent transformants of each allele (FIGURE 4.3) – presumably due to TDNA positional and copy number effects. However, when the distributions and means of each pool of transformants are compared, several striking observations emerge. First, the phenotype distributions for the Col and L*er* alleles are very different. Specifically, the Col allele recovers the mutant trichome density phenotype much better than the L*er* allele. None of the transformants carrying the L*er* allele had phenotypes beyond the range typically observed for untransformed *atmyc1* mutant plants. By comparison, independent transformants with the Col allele show phenotypes ranging from *atmyc1*-like (relatively few hairs) to overexpressor-like (extremely hairy). If we consider the average effect of each allele, the findings are very much in accord with the outcome of the quantitative complementation tests reported above. In both cases, the L*er* allele was completely unable to recover the mutant phenotype and the Col allele completely restored trichome initiation back to Col wildtype levels. A Kruskal-Wallis test showed that there are significant differences in the

phenotypic effects of different alleles and subsequent Tukey's HSD tests revealed that the only pairwise significant differences in phenotype are between the pool of L*er* transformants and each of the pooled data for each of the other alleles (TABLE 4.1). This observation reveals two other major points. First, L*er* and CVI, the two parents of another population in which TDL5 was mapped, possess alleles with significantly differing transgenic trichome number phenotypes. Second, the Sha allele, which has high sequence similarity to the L*er* allele, recovers the mutant phenotype at least as well as the Col and CVI alleles.

Cluster analyses of *ATMYC1* alleles

Considerable sequence variation was detected among natural alleles of *ATMYC1* among 76 accessions. Phylogenetic analyses yielded a gene tree that is essentially split into two very strongly supported clusters (FIGURE 4.4); to ease discussion, alleles in each cluster are called either Type I or Type II. Curiously, both allele types occur at very high frequency; of the 76 accessions for which *ATMYC1* was sequenced, 33 possess a Type I allele and 43 have a Type II allele. Several smaller clusters nested within the two major types are also well-supported. With regard to the populations in which TDL5 was mapped, it is interesting that the six parental accessions fall into five different subtypes. Perhaps most interesting among these subtypes is cluster IIa2, in which the L*er* allele falls. Only three other accessions are in this cluster, one of which is La-0, the wildtype (non-*er*) Landsberg accession.

Molecular evolution of the *ATMYC1* locus

An analysis of the 76 *A. thaliana* alleles of *ATMYC1* yielded overall levels of nucleotide diversity (π & θ_w; TABLE 4.2) that are comparable to values reported for other *A. thaliana* genes (reviewed in (KAMIYA et al. 2002)). However, a sliding window analysis revealed very high localized levels of nucleotide diversity (FIGURE 4.5). Although most of the diversity peaks overlie introns, the highest level of diversity was detected within exon six.

Because regions of high nucleotide diversity seemed to correspond with divergence between Type I and Type II alleles, we wanted to characterize the nature of this molecular variation. Specifically, we wanted to test whether the nucleotide variation that differentiates the two allele types is neutral or potentially functional. To explore this, we used a sliding window method to study rates of non-synonymous (K_A) and synonymous (K_S) divergence across the entire 1.56 kb coding region between Types I and II alleles. These analyses revealed evidence of alternative forms of selection that are region-specific (FIGURE 4.6). Across most of the gene, it appears that purifying selection has acted to constrain the amino acid sequence (ratios $<<1$), however within exon six, extremely high rates of amino acid replacement have taken place. As a K_A/K_S ratio greater than one is often cited as a very conservative cut-off for positive selection (KIMURA 1977; PRESGRAVES et al. 2003), values approaching 20 are exceptional. Even the more rigorous approach of using the gene-wide average K_S value still resulted in values greater than five. Outside of exon six, the only other region that well exceeds the cut-off for positive selection is the extreme 5' end of exon two.

To compare levels of gene divergence that have taken place within *A. thaliana* to divergence that has taken place since *A. thaliana* diverged from related taxa, we calculated the overall K_A/K_S ratios (non-sliding window analyses) and plotted them onto the gene tree (FIGURE 4.4) at the branching points that define the splits between *A. thaliana* and all four outgroups and between the two *A. thaliana* allele types. The ratio for between *A. thaliana* types is more than twice that of the ratio for between species. In short, this means that more than twice as much molecular divergence has taken place within *A. thaliana* than has taken place between *A. thaliana* and four outgroup species.

DISCUSSION

ATMYC1 is a QTL (TDL5) for trichome density

ATMYC1 was initially identified as a candidate locus for the QTL, TDL5, based on its physical position (FIGURE 4.1) and high sequence similarity with two other bHLH loci (HEIM et al. 2003; TOLEDO-ORTIZ et al. 2003) known to have trichome density phenotypes upon knock-out (PAYNE et al. 2000; ZHANG et al. 2003). We determined that a TDNA insertion line of *ATMYC1* has a reduced trichome density phenotype and subsequently used this mutant to perform two forms of quantitative complementation tests. Both genetic and transgenic complementation data show that alternative alleles of *ATMYC1* vary in their ability to recover the *atmyc1* mutant phenotype. Because the *atmyc1* mutant is in the Col-0 background, the most direct comparisons that could be made with regard to results from previous mapping studies were between the Col-0 and L*er*-0 alleles – Col was a parent only in the L*er* x Col mapping population. These

specific comparisons revealed that the Col allele completely recovers the *atmyc1* trichome density phenotype back to wildtype levels (crossing data) and beyond (complementation data), while the L*er* allele was unable to jockey the phenotype beyond *atmyc1* mutant levels in both sets of complementation tests. Whether the phenotype difference is due to amino acid substitutions, unobserved sequence divergence in promoter regions, or some combination of these remains unknown. However, it is clear that the *ATMYC1* locus is a QTL for trichome density. Given the physical position estimated for TDL5 and the large additive effect of substituting Col alleles for L*er* alleles, it is most likely that *ATMYC1* is the TDL5 QTL mapped in the Col x L*er* population (SYMONDS et al. in press).

Because of positional overlap among QTL mapped to the top of chromosome four in different populations (FIGURE 4.1), it was suspected that the same locus was mapped independently four times. Three of those populations had the L*er* accession as one of the parents: L*er* x Col, L*er* x CVI, and L*er* x No-0. Because in complementation tests the L*er* allele behaved like a null, it seemed very likely that the reduced trichome density phenotype caused by this allele was also underlying the QTL mapped at the top of chromosome four in the L*er* x CVI and L*er* x No-0 populations. Indeed, our transgenic results show that the CVI allele recovers the *atmyc1* mutant phenotype just as well as the Col allele and much better than the L*er* allele. So, for both the L*er* x Col and L*er* x CVI populations, we have strong evidence that variation at *ATMYC1* is responsible for the mapping of TDL5. Due to large insertions immediately flanking the open reading frame of *ATMYC1* genomic alleles from the No-0 and Bay accessions, cloning has proven

difficult. So, any potential difference between the L*er* and No-0 alleles and the Bay and Sha alleles are as yet untested in transgenic plants.

Cluster analyses reveal two well-supported allele types for *ATMYC1*

The identification of *ATMYC1* as a QTL for trichome density prompted a survey of allelic variation and allele frequencies among ecotypes of *A. thaliana*. This survey revealed striking levels of intraspecific variation. Cluster analyses showed strong support for two high-frequency allele types, Types I and II (FIGURE 4.4). This sort of allelic dimorphism has been reported for some, but not all, other loci in *A. thaliana* (INNAN et al. 1996; KAWABE et al. 1997; TIAN et al. 2002). The source of this bifurcating pattern remains undetermined, but the diversity discovered has been hypothesized to be maintained by balancing selection or is a result of relatively recent demographic events, such as the fusion of two long-isolated populations.

Our transgenic data suggest that some members of Type I and Type II alleles may perform the role of *ATMYC1* in trichome initiation equally well. If this is true, then the maintenance of these two types is unlikely to be the result of balancing selection. Alternatively, balancing selection may be acting on loci linked to Atmcy1, and variation at this locus is maintained simply through linkage. However, a preliminary screen identified no obvious recombinant alleles of *ATMYC1*. Yet another alternative is that balancing selection has acted on phenotypes pleiotropic to trichome density. Recent work has shown that *ATMYC1* interacts with anthocyanin pathway-specific proteins in vitro (ZIMMERMANN et al. 2004), and paralogs of *ATMYC1* are known to have pleiotropic mutant phenotypes (ZHANG et al. 2003).

A phylogram derived from inferred cDNA sequences shows that most of *ATMYC1* sequence diversity within the species falls along the branch that separates the two major types and comparatively less diversity resides within each type. However, there is sufficient within-type variation to distinguish among several well-supported subtypes. Interestingly, the six parent accessions used in prior mapping experiments are of five different subtypes (FIGURE 4.4). Notably, the L*er* allele is shared with only three other ecotypes (subtype IIb-2), one of which is the wildtype (non-*er*) Landsberg accession. The fact that L*er* has an allele that is identical to the wildtype Landsberg line is most interesting because it suggests that the null allele carried by L*er* is of natural origin, as opposed to being a by-product of mutagenesis, as is the *erecta* mutation (TORII et al. 1996). This is supported by the fact that La-0 has an identical trichome density phenotype to L*er* (personal observation).

Types I and II *ATMYC1* alleles demonstrate strongly divergent molecular evolution

Although comparisons of the genomic sequences of *ATMYC1* alleles showed that the bulk of molecular variation is within introns, the highest regional levels reside within an exon. Both nucleotide diversity and polymorphism measures (π and θ) from across all *A. thaliana* genomic alleles and levels of amino acid replacement (K_A/K_S ratios) between Type I and Type II alleles were highest in exon six (FIGURES 4.5 & 4.6). While the majority of the coding region yielded K_A/K_S ratios much less than one, which is indicative of purifying selection, several sections of exon six produced ratios much greater than one, suggesting that positive selection has acted on this region. Whether or not the extreme spikes observed when using local K_S values are accurate, both the local

K_S and more conservative gene-wide K_S methods yielded ratios significantly greater than one between positions 700 and 892 bp within exon six. Within this span of 64 amino acids, there are 15 fixed amino acid replacements (23%) between Type I and Type II alleles (FIGURE 4.7). Again, this is very strong evidence that positive selection has acted to diverge these two high-frequency allele types.

What has driven divergence between the two types of *ATMYC1* alleles is unknown. However, variation observed between *A. thaliana* alleles and outgroup alleles are suggestive of what began the process. Within exon six, two indels are present that distinguish all *A. thaliana* and *A. lyrata* alleles from the remaining three outgroups, *Arabis griffithiana*, *Crucihimalaya himalaica*, and *Capsella bursa-pastoris* (FIGURE 4.7). Given the conservation of these indels within the outgroups, the most parsimonious scenario is that each is due to a single deletion event along the branch leading to *A. thaliana* and *A. lyrata*. The alternative, that there were multiple independent insertions (of the same sequence) in each of the non-monophyletic outgroup species is untenable. The deletion at ~position 730 is 18 bp (6 amino acids) and the deletion at position ~870 is 15 bp (5 AA), so the reading frame has either been maintained or recovered through subsequent, presumably, small indels. Most curiously, these two indels flank the region of exon six that now demonstrates such high levels of nucleotide variation within *A. thaliana* (FIGURES 4.5, 4.6, and 4.7). It seems plausible then that these two deletions initially caused dramatic changes in protein function - perhaps under conditions where the resulting phenotype changes were inconsequential - and subsequent changes in selection pressure fixed new mutations that regained, restored, or bettered protein function - twice. So, rather than the K_A/K_S-inferred positive selection acting to

differentiate the two allele types found within *A. thaliana*, it seems that positive selection may have acted on each allele lineage independently to evolve away from a suboptimal ancestral allele. If this is the case, then it appears to have been successful, as representatives of each allele type were able to recover the *atmyc1* mutant phenotype equally well. The fact that L*er* possesses a Type II allele that is non- or hypo-functional for trichome density may say nothing of Type II alleles as a whole, but rather may reflect low frequency mutations (Type IIa_2 alleles only) in the coding region or as yet unobserved mutations in promoter regions that affect gene regulation.

Conclusions

(1) As transgenes, the genomic alleles of *ATMYC1* from Col, CVI, and Sha are functionally equivalent, while the L*er* allele behaves like a nonfunctioning null. (2) Cluster analyses revealed that the Bay and No-0 accessions possess identical *ATMYC1* alleles. (3) Mapping data have shown that for TDL5, Col, CVI, and No-0 have alleles that are functionally superior to L*er*, and the Sha allele outperforms the Bay allele. (4) Complementation tests indicate that *ATMYC1* is TDL5. Consideration of all of these points lead us to tentatively propose the following hierarchy of *ATMYC1* allele function for trichome density:

(Col = CVI = Sha) > (Bay = No-0) > L*er*

In terms of gene tree allele types, this can be rewritten as:

(Ia = Ib = IIb) > IIa_1 > IIa_2

These models provide a set of testable hypotheses that will motivate future research.

Tables

Table 2.1 List of the 126 accessions used in this study

Stock #	Background	Stock #	Background	Stock #	Background
1394	No-0	6798	Mir-0	6843	Pt-0
1516	Sf-2	6799	Mt-0	6844	Ra-0
3081	No-0	6800	Mz-0	6845	Rd-0
6173	Est	6801	Na-1	6848	Rsch-0
6175	Condara	6803	Nd-0	6850	Rsch-4
6600	Aa-0	6804	Nie-0	6851	Ru-0
6601	Ag-0	6805	No-0	6852	Se-0
6607	Ba-1	6806	Np-0	6853	Sei-0
6608	Bay-0	6807	Nok-0	6854	Sap-0
6615	Bl-1	6808	Nok-1	6855	Sf-1
6616	Bla-1	6809	Nok-2	6856	Sav-0
6624	Bla-12	6810	Nok-3	6857	Sf-2
6625	Bla-14	6811	Nw-0	6859	Sg-2
6626	Br-0	6812	Nw-1	6860	Sh-0
6627	Bs-1	6813	Nw-2	6861	Si-0
6643	Bir-0	6814	Nw-3	6862	Sp-0
6659	Cal-0	6815	Nw-4	6863	St-0
6660	Can-0	6816	Ob-0	6864	Ste-0
6664	Chi-0	6818	Ob-2	6865	Stw-0
6665	Chi-1	6819	Ob-3	6867	Ta-0
6666	Chi-2	6820	Old-1	6868	Ts-1
6669	Co-1	6821	Old-2	6869	Ts-2
6672	Co-4	6822	Or-0	6870	Ts-3

Table 2.1 continued

Stock #	Background	Stock #	Background	Stock #	Background
6675	Cvi-0	6823	Ove-0	6871	Ts-5
6685	Dra-0	6824	Oy-0	6872	Ts-6
6686	Dra-1	6825	Pa-1	6873	Ts-7
6700	Est-0	6826	Pa-2	6874	Tsu-0
6701	Est-1	6827	Pa-3	6878	Ty-0
6702	Et-0	6828	Per-1	6879	Uk-1
6716	Gd-1	6829	Per-2	6880	Uk-3
6723	Gr-1	6830	Per-3	6884	Van-0
6734	Hau-0	6831	Pf-0	6885	Wa-1
6736	Hi-0	6832	Pi-0	6889	Wil-2
6745	Jl-3	6833	Pi-2	6891	Ws-0
6751	Kas-1	6834	Pla-0	6897	Wu-0
6754	Kil-0	6835	Pla-1	6898	X-0
6762	Kn-0	6836	Pla-3	6899	XX-0
6765	La-0	6838	Pn-0	Cal-0	Calvert
6769	Lc-0	6839	Po-0	Col	Col-0
6783	Ll-2	6840	Po-1	Ler	Ler-0
6784	Lm-2	6841	Pr-0	Tac-0	Tacoma
6789	Ma-0	6842	Pog-0	WS	WS

Table 2.2 Microsatellite locus table

Marker	Chr	cM	Repeat Type	% Amp.	Allele Range	# of Alleles	Gene Diversity	SMM[b]	TPM[c]	IAM[d]
nga59	1	1.6	CT	92	111-192	31	0.94	*	ns	ns
ZFPG	1	37.4	TC	98	127-236	26	0.87	**	*	ns
Centromere	1	62.0	-	-	-	-	-	-	-	-
nga128[a]	1	83.0	TC	90	177-227	13	0.90	ns	ns	ns
nga692	1	119.0	GA	83	106-152	25	0.90	**	*	ns
nga1145[a]	2	9.6	GA	94	208-239	11	0.45	****	****	***
Centromere	2	19.0	-	-	-	-	-	-	-	-
nga1126	2	50.6	GA	93	182-221	17	0.87	*	ns	ns
AthUbique[a]	2	82.0	CT	95	164-172	5	0.52	ns	ns	ns
nga172[a]	3	7.0	GA	92	152-244	31	0.95	ns	ns	ns
GaPAB	3	43.8	TTC	98	135-150	4	0.50	ns	ns	ns
Centromere	3	59.0	-	-	-	-	-			
nga707	3	78.0	TC	83	119-141	10	0.52	****	**	*
CIW5[a]	4	5.3	TA	77	155-200	14	0.62	****	****	*
Centromere	4	20.0	-	-	-	-	-	-	-	-
nga8	4	24.2	GA	91	122-222	38	0.96	ns	ns	ns
CIW7[a]	4	65.0	TA	87	126-180	22	0.92	*	ns	ns
nga1139	4	83.4	TC	97	74-142	22	0.93	ns	ns	*
nga1107	4	105.0	GA	90	134-155	9	0.41	****	***	*
nga249	5	23.1	TC	98	115-139	11	0.49	****	****	*
CDPK9	5	44.5	TC	87	86-179	22	0.86	***	**	ns
Centromere	5	70.0	-	-	-	-	-	-	-	-
CIW9[a]	5	88.0	TA	83	138-208	28	0.89	***	***	*
nga129[a]	5	105.0	GA	83	179-205	14	0.71	****	****	ns
CIW10[a]	5	128.0	TA	83	136-187	20	0.93	ns	ns	*

* $p<0.05$
*** $p<0.01$
*** $p<0.005$
**** $p<0.005$

[a] Loci utilizing the M13 tailing scheme. True allele sizes for these loci are expected to be 19 bp smaller than those reported here because of the use of the M13 forward sequence.
[b] Goodness of fit test for stepwise mutation model
[c] Goodness of fit test for two parameter model
[d] Goodness of fit test for infinite alleles model

Table 2.3 Correlation coefficients for microsatellite classes and diversity

	All[a]		Low d[b]		High d[c]	
Repeat	U[d]	T[e]	U	T	U	T
Test						
GA/TC						
Pearson's	0.78***	0.70**	0.86*	0.29	0.84**	0.83**
Spearman's	0.93***	0.86***	0.81	-0.06	0.76*	0.76*
TA/AT						
Pearson's	0.89	-0.63	-	-	-	-
Spearman's	0.95	-0.20	-	-	-	-

[a] All loci of a particular repeat type
[b] Low diversity loci only (d<0.75)
[c] High diversity loci only (d>0.75)
[d] Largest number of uninterrupted repeats at a locus
[e] Total number of repeats at a locus
* $p<0.05$
** $p<0.01$
*** $p<0.001$

Table 2.4 Frequency of mutation model test rejection

	Test			
Locus type	SMM	TPM	IAM	Average
Low diversity*	0.86	0.86	0.71	0.81
Hgih diversity	0.58	0.33	0.25	0.37

*Excluding the trinucleotide locus, GapAB

Table 3.1 Mapping population details

Mapping population	RILs[a]	Markers[b]	H^{2c}	QTL[d]	Linkage group[e]				
					1	2	3	4	5
L*er*-0 x Col-4	92	125	0.795	3	114	68	71	82	100
L*er*-2 x CVI	103	242	0.896	5	127	82	83	83	115
L*er*-2 x No-0	96	46	0.877	4	70	88	93	67	83
Bay x Shah	144	38	0.892	3	85	63	68	72	76

[a] Number of recombinant inbred lines used
[b] Number of markers used in a given population
[c] Broad-sense heritability estimate
[d] Number of QTLs identified
[e] Linkage group lengths in cM for chromosomes 1-5

Table 3.2 Trichome density QTL details

		Mapping Population													
		L*er* x Col				L*er* x CVI				L*er* x No-0				Bay x Sha				
Ch[a]	QTL[b]	Pos[c]	PE[d]	A[e]	L*er*[f]	Pos	PE	A	L*er*	Pos	PE	A	L*er*	Pos	PE	A	Bay[f]	Cand[g]
1	TDL1									36.2	11.6	3.0	↓					
2	TDL2	51.6	21.0	4.6	↓	**68.4**	19.4	5.2	↓	63.2	11.5	2.8	↓	46.6	12.7	4.2	↓	TTG2
3	TDL3	36.1	22.1	5.6	↓													GL1
3	TDL4													51.0	11.8	4.1	↓	
4	TDL5	3.5	9.1	3.0	↓	2.0	12.1	3.0	↓	11.0	11.4	3.0	↓	1.0	25.0	6.0	↓	
4	TDL6					**48.5**	16.5	4.1	↑									
4	TDL7									60.2	11.0	2.9	↑					
5	TDL8					32.4	7.0	2.6	↑									TTG1
5	TDL9					68.9	6.8	2.4	↓									GL3

QTL positions in bold indicate loci that epistatically interact
[a] Chromosome on which QTL was mapped
[b] QTL identifier
[c] Map position in cM
[d] Percent of variation within mapping population explained by the QTL
[e] Additive effect of QTL (2a)
[f] Direction of effect on trichome density of either the L*er* or Bay allele (depending on the population) for each QTL
[g] Candidate gene based on physical proximity to QTL position

Table 4.1. Complementation data statistics

Test	Test statistic	Significance threshold	Significance
Kruskal-Wallis	H = 14.36	$\chi^2_{.005[3]} = 12.84$	***
Tukey's HSD			
Col-L*er*			**
Col-CVI			NS
Col-Sha			NS
CVI-Sha			NS
CVI-*Ler*			**
Sha-L*er*			**

** <0.01
*** <0.005

Table 4.2 Indices of genetic diversity

Taxa	n^a	π^b	$\theta_w{}^c$	$K_A/K_S{}^d$
All taxa	80	0.01582 ± 0.00261	0.03341 ± 0.00858	0.243
A. thaliana	76	0.01061 ± 0.00038	0.00878 ± 0.00244	0.243
Type I only	33	0.00236 ± 0.00036	0.00405 ± 0.00143	0.253
Type II only	43	0.00133 ± 0.00014	0.00278 ± 0.00099	0.234

[a] number of samples

[b] nucleotide diversity – average proportion of nucleotide differences between all possible pairs

[c] nucleotide polymorphism – proportion of nucleotide sites expected to be polymorphic

[d] K_A/K_S within each category

Figures

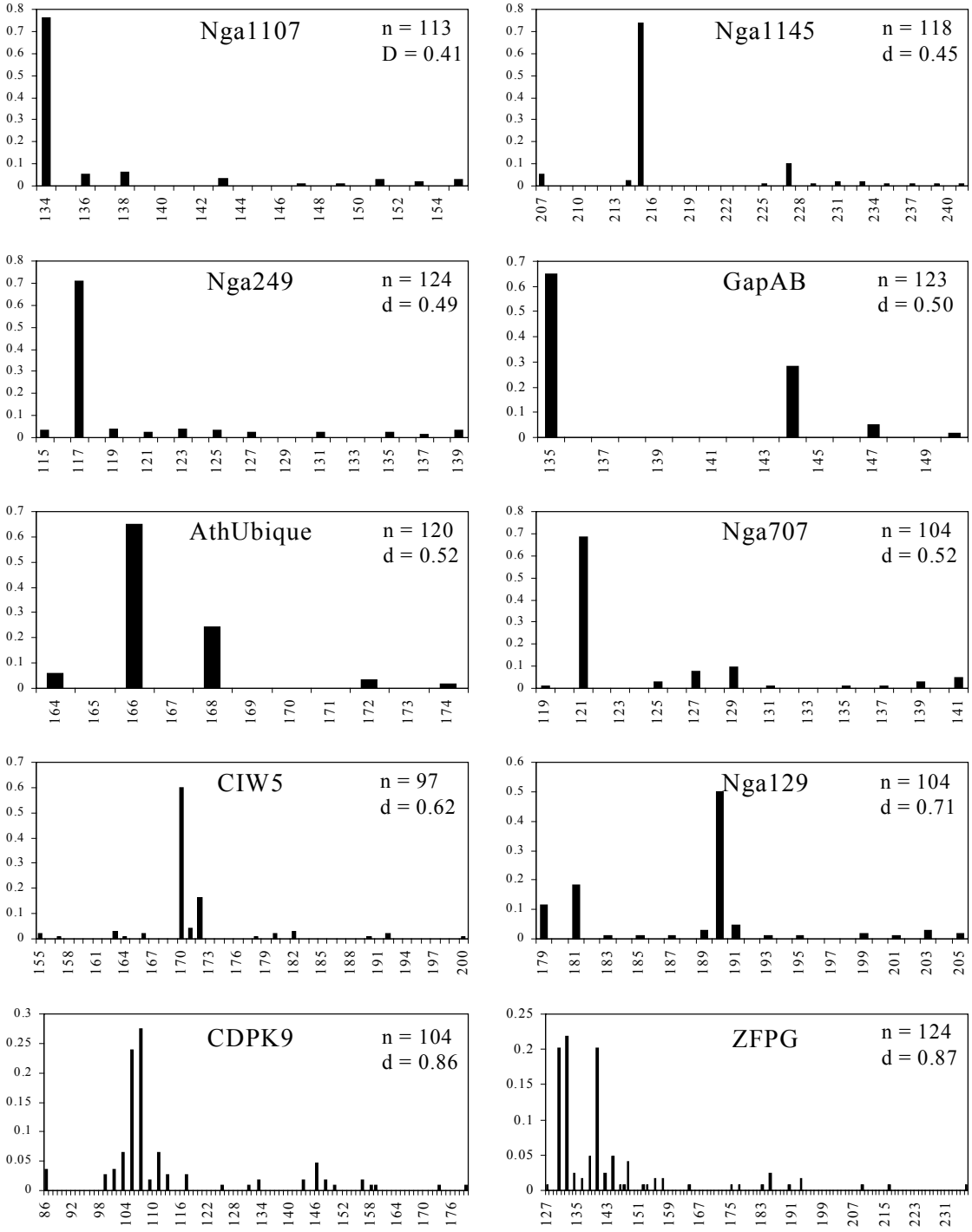
Nga1107
n = 113
D = 0.41
Nga1145
n = 118
d = 0.45
Nga249
n = 124
d = 0.49
GapAB
n = 123
d = 0.50
AthUbique
n = 120
d = 0.52
Nga707
n = 104
d = 0.52
CIW5
n = 97
d = 0.62
Nga129
n = 104
d = 0.71
CDPK9
n = 104
d = 0.86
ZFPG
n = 124
d = 0.87

Figure 2.1

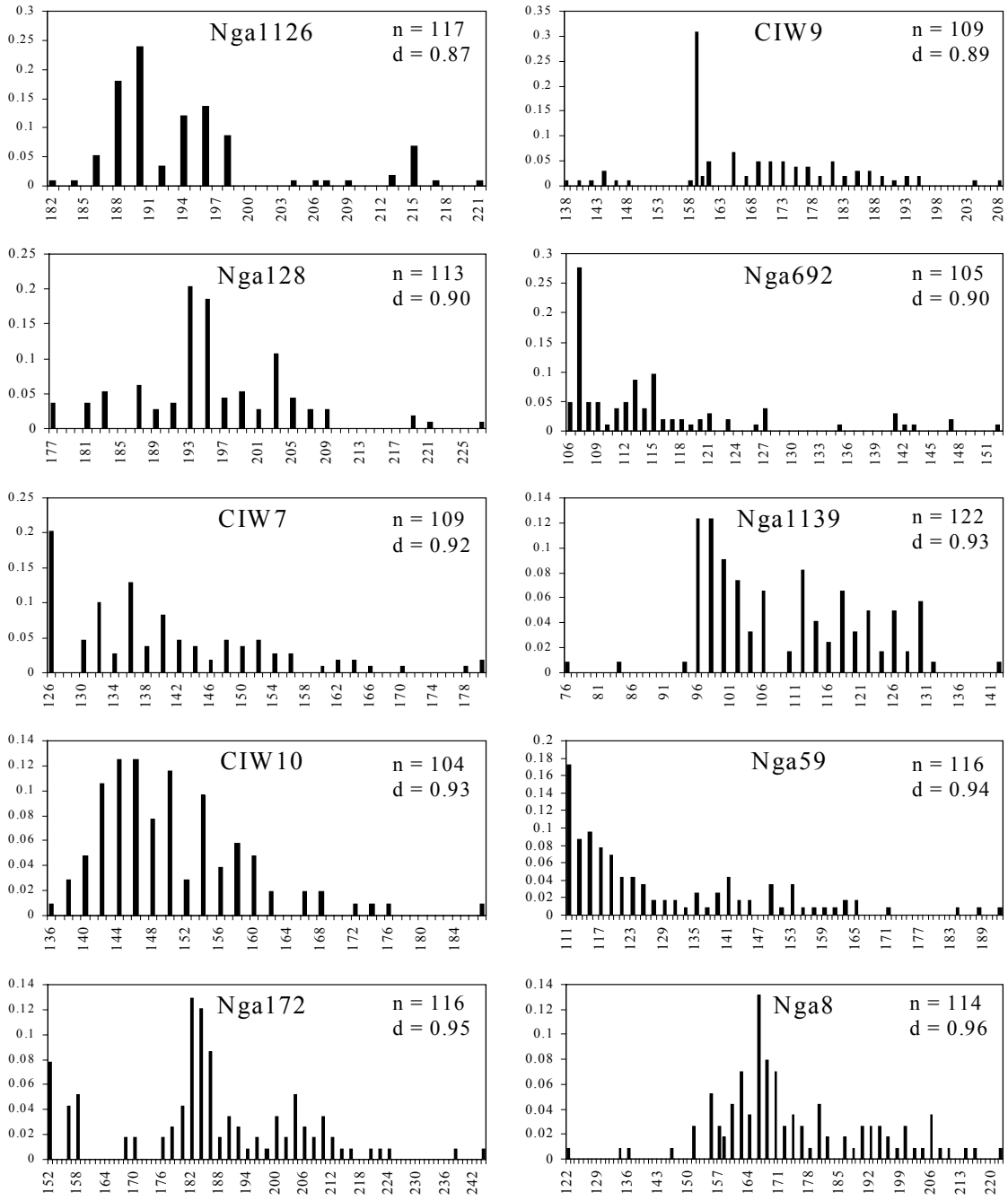

Figure 2.1 Histograms showing allelic distributions for the 20 microsatellite loci examined. Allele size in bp is shown along the x-axis of each chart and the frequency of each allele class is displayed along the y-axis. Loci are arranged from lowest to highest gene diversity. Sample size (n) and gene diversity (d) are shown in the upper right of each chart.

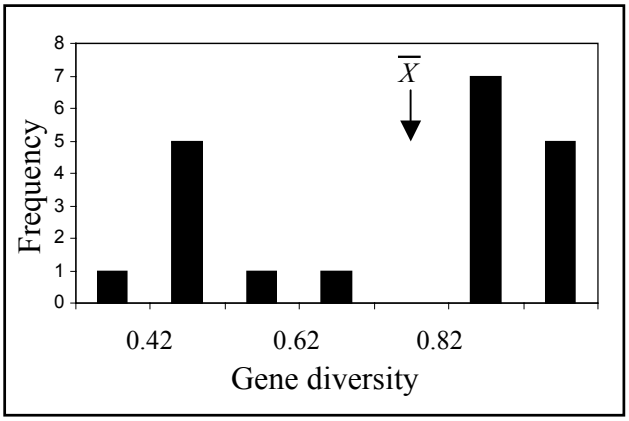

Figure 2.2 Distribution of gene diversity measures among the 20 microsatellite loci. $\overline{X}$ marks placement of the mean.

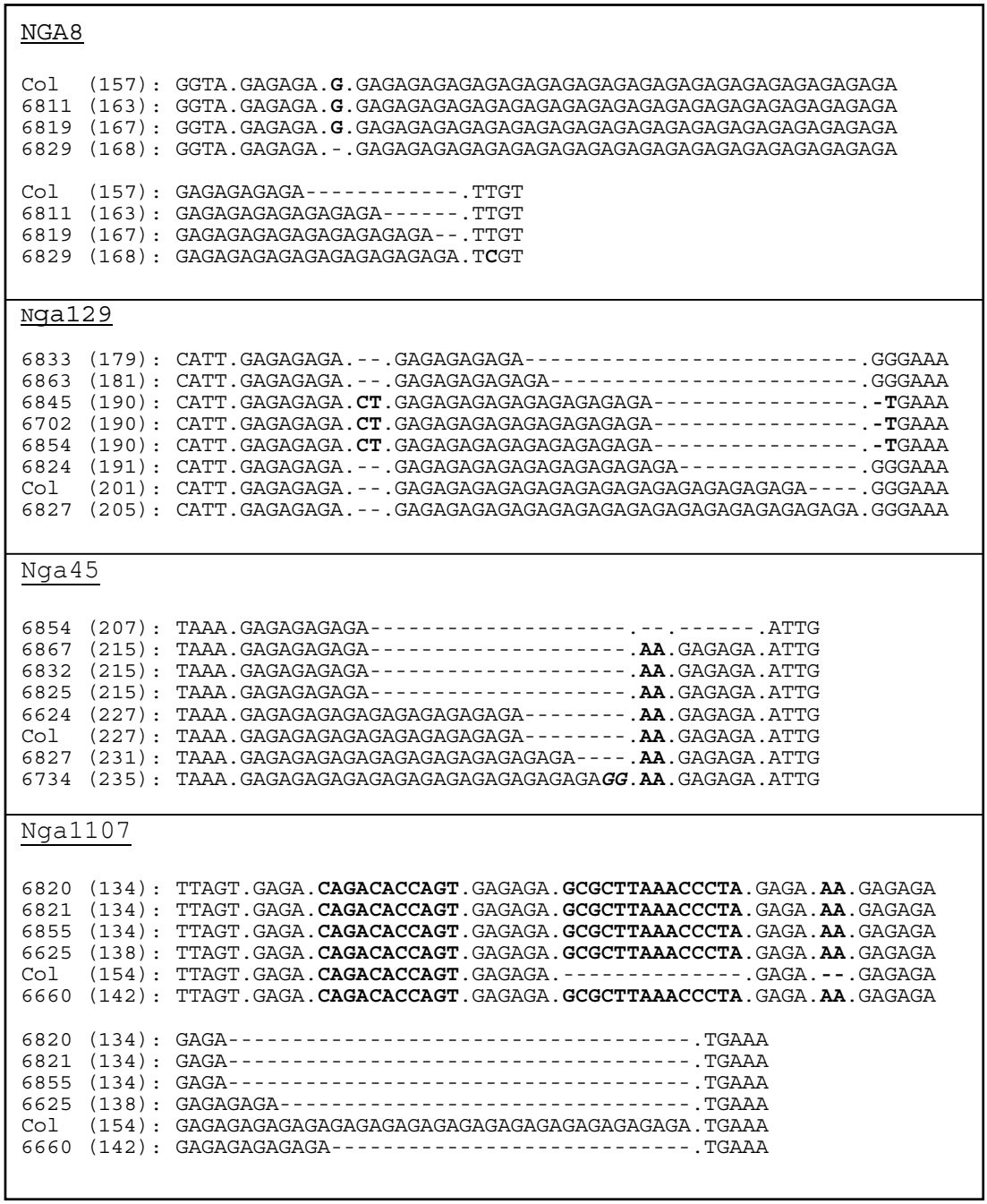

Figure 2.3 Sequence alignments for representative high (nga8) and low (nga129, nga1145, nga1107) gene diversity loci. Interruptions within repeat regions are highlighted in bold. As the only source of variation revealed at two of the high diversity loci sequenced, CIW7 and nga172, was repeat number variation, only sequences from a representative locus, nga8 are shown. "-" indicate gaps and "." serve to break up the sequence to aid viewing.

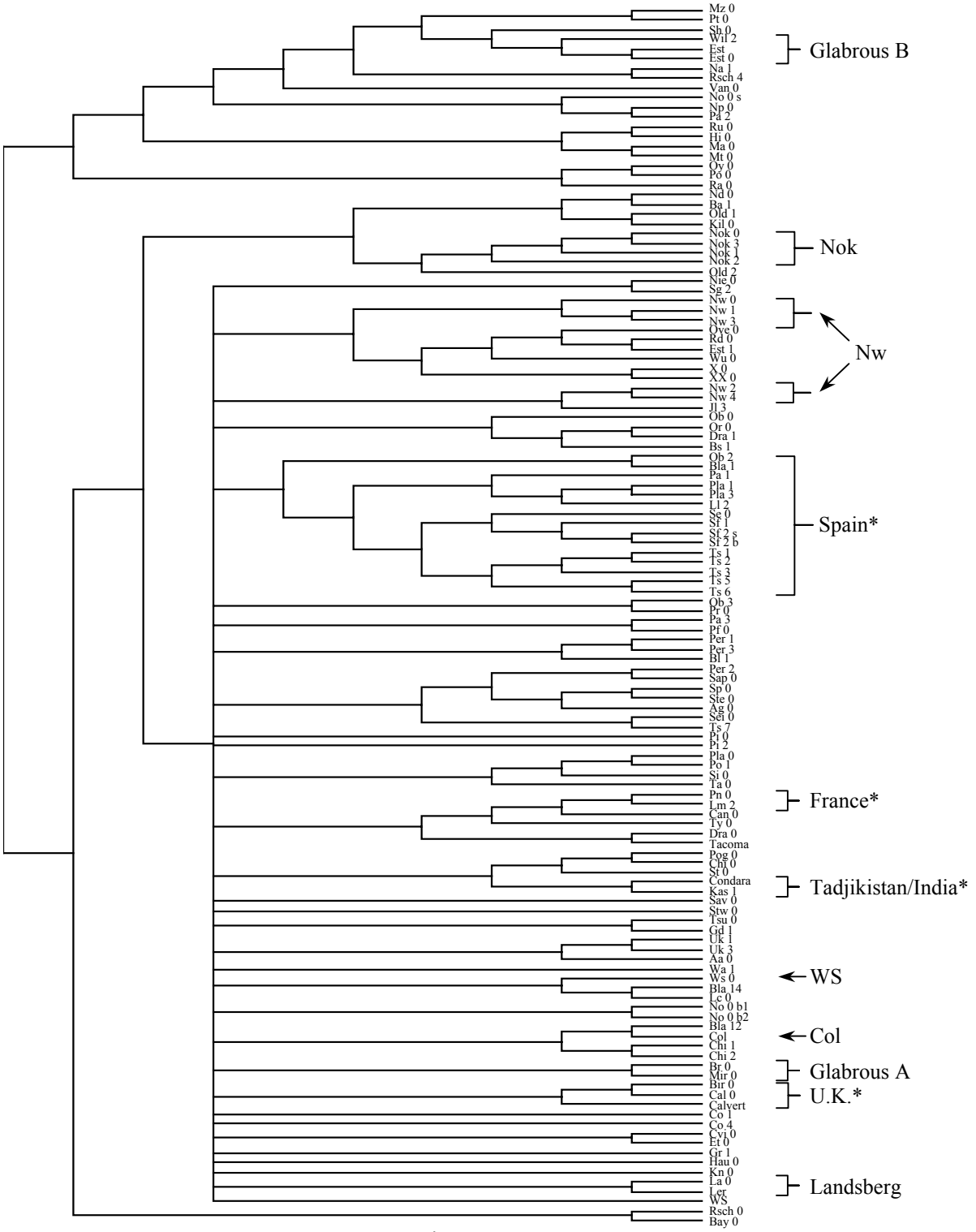
Mz 0
Pt 0
Sh 0
Wil 2
Est
Est 0
Na 1
Rsch 4
Van 0
No 0 s
Np 0
Pd 2
Ru 0
Hi 0
Ma 0
Mt 0
Oy 0
Po 0
Ra 0
Nd 0
Ba 1
Old 1
Kil 0
Nok 0
Nok 3
Nok 1
Nok 2
Old 2
Nie 0
Sg 2
Nw 0
Nw 1
Nw 3
Ove 0
Rd 0
Est 1
Wu 0
Y 0
XX 0
Nw 2
Nw 4
Jl 3
Ob 0
Or 0
Dra 1
Bs 1
Ob 2
Bla 1
Pa 1
Pla 1
Pla 3
Ll 2
Se 0
Sf 1
Sf 2 s
Sf 2 b
Ts 1
Ts 2
Ts 3
Ts 5
Ts 6
Ob 3
Pr 0
Pa 3
Pf 0
Per 1
Per 3
Bl 1
Per 2
Sap 0
Sp 0
Ste 0
Ag 0
Sei 0
Ts 7
Pi 0
Pi 2
Pla 0
Po 1
Si 0
Ta 0
Pn 0
Lm 2
Can 0
Ty 0
Dra 0
Tacoma
Pog 0
Chi 0
St 0
Condara
Kas 1
Sav 0
Stw 0
Tsu 0
Gd 1
Uk 1
Uk 3
Aa 0
Wa 1
Ws 0
Bla 14
Lc 0
No 0 b1
No 0 b2
Bla 12
Col
Chi 1
Chi 2
Br 0
Mir 0
Bir 0
Cal 0
Calvert
Co 1
Co 4
Cvi 0
Et 0
Gr 1
Hau 0
Kn 0
La 0
Ler
WS
Rsch 0
Bay 0
Glabrous B
Nok
Nw
Spain*
France*
Tadjikistan/India*
WS
Col
Glabrous A
U.K.*
Landsberg

Figure 2.4

Figure 2.4 Majority-rule (70%) consensus tree derived from 1000 independent UPGMA runs. Clusters and accessions of particular interest and that are discussed in the text are denoted by brackets. Examples of accessions that cluster together according to geographic origin are marked with an "*". For details on UPGMA analyses, see Materials and Methods.

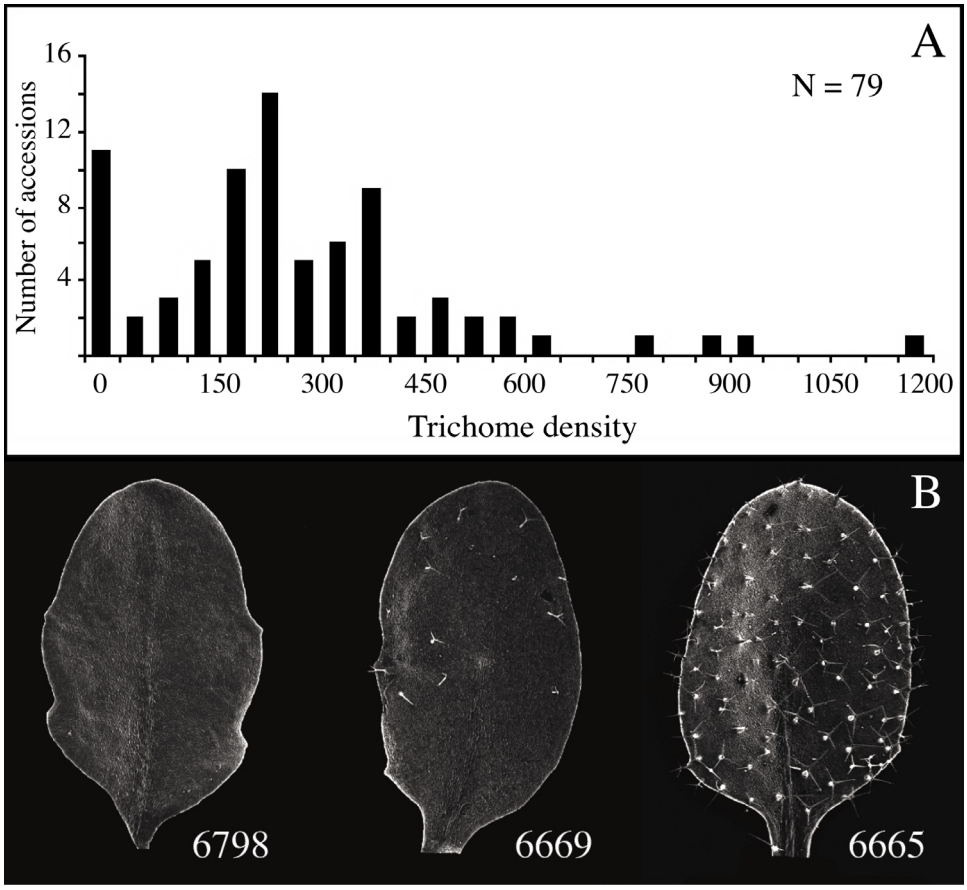

Figure 3.1 Trichome density variation of the third true leaf. (A) Distribution of trichome density on the third true leaf for 79 natural accessions of *A. thaliana*. (B) Scanning electron micrographs of the adaxial surface of the third true leaf from three natural accessions of *A. thaliana* that represent the range of trichome density, from completely glabrous to densely hairy. The numbers below each leaf refer to ABRC accession numbers.

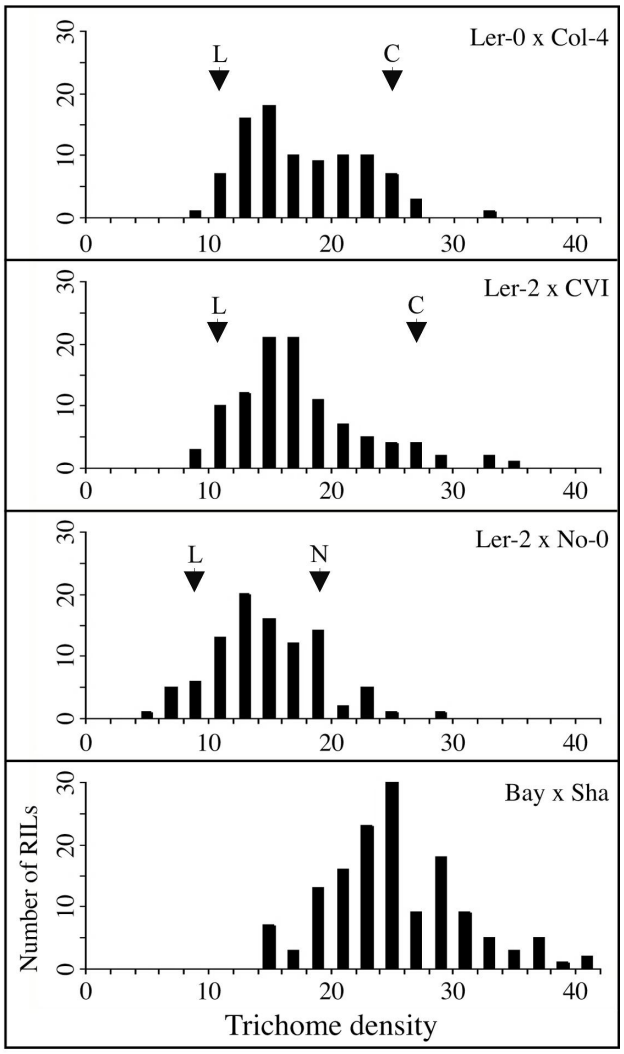

Figure 3.2 Trichome density phenotype distributions for RILs of each mapping population. Letters above graphs indicate the mean phenotype for the two parents of each population.

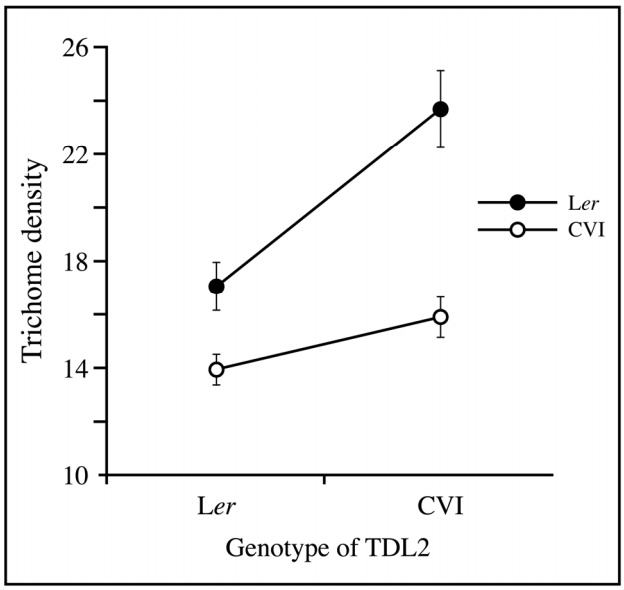

Figure 3.3 QTL with significant epistatic interactions in the L*er*-2 x CVI population. Plotted points represent two-locus genotype means ±SE for TDL2 X TDL6.

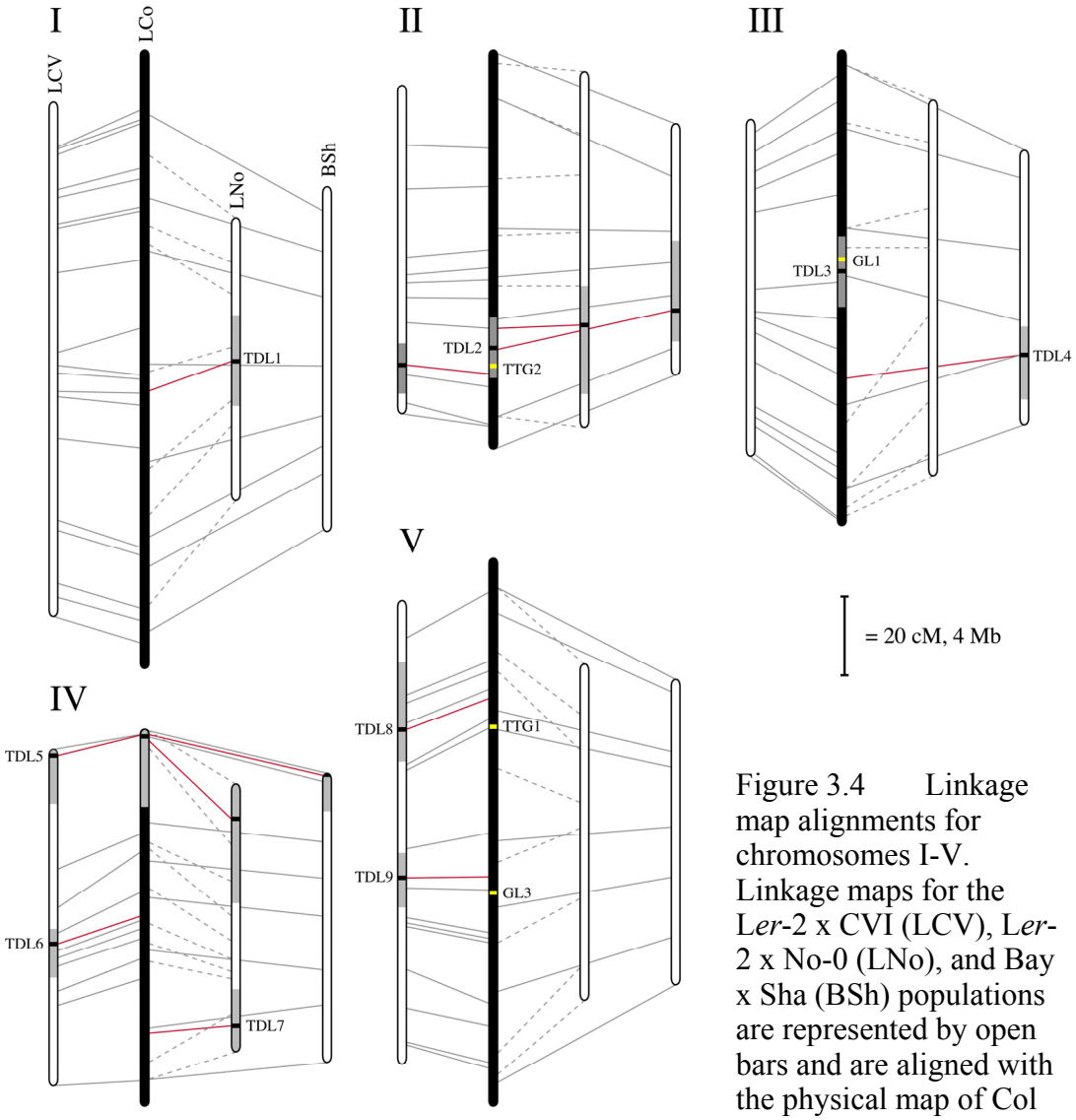

Figure 3.4 Linkage map alignments for chromosomes I-V. Linkage maps for the L*er*-2 x CVI (LCV), L*er*-2 x No-0 (LNo), and Bay x Sha (BSh) populations are represented by open bars and are aligned with the physical map of Col (solid bar). Small, black horizontal lines within bars indicate QTL positions and the grey area around each QTL represents the 2-LOD support interval; QTL mapped in the L*er*-0 x Col-4 population are shown on the Col physical map. The linkage positions of markers used in each population are aligned with their corresponsing physical position on the Col map (gray lines). Dashed lines were used to distinguish lines connecting markers to the L*er*-2 x No-0 population from those connecting to the Bay x Shah population. To compare the positions of QTL mapped in different populations, each was connected to its estimated physical position on the physical map (red lines). The positions of candidate genes are shown in yellow, with their corresponding gene name at the right.

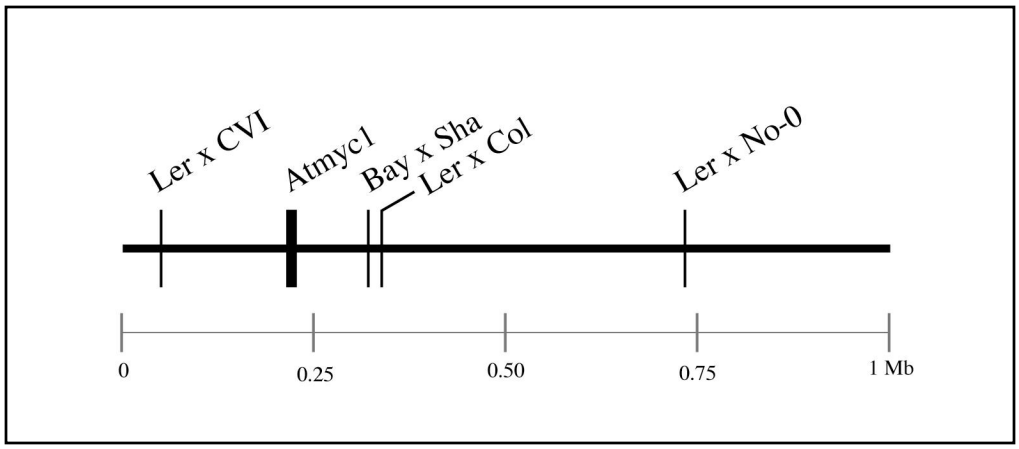

Figure 4.1 Schematic of the first 1 Mb of the physical map at the top of chromosome four. The physical position of the *ATMYC1* locus and the estimated physical positions of TDL5 from four different mapping populations are indicated to scale.

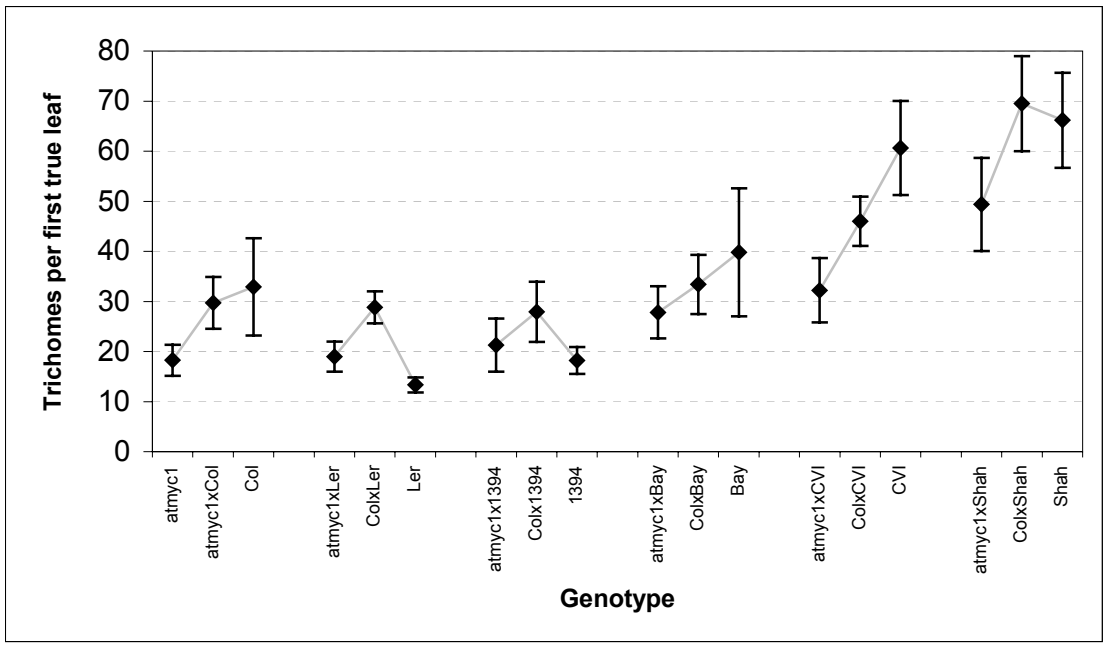

Figure 4.2 Complementation test results. Plotted are the mean number of hairs on first true leaves for the *atmyc1* mutant, parent accessions, and F1s from crosses between parent accessions and Col-0 and *atmyc1*. Gray lines connect genotype comparisons and error bars reflect standard deviations of each mean. Note that all crosses made with *atmyc1* have lower trichome numbers relative to crosses made with Col-0.

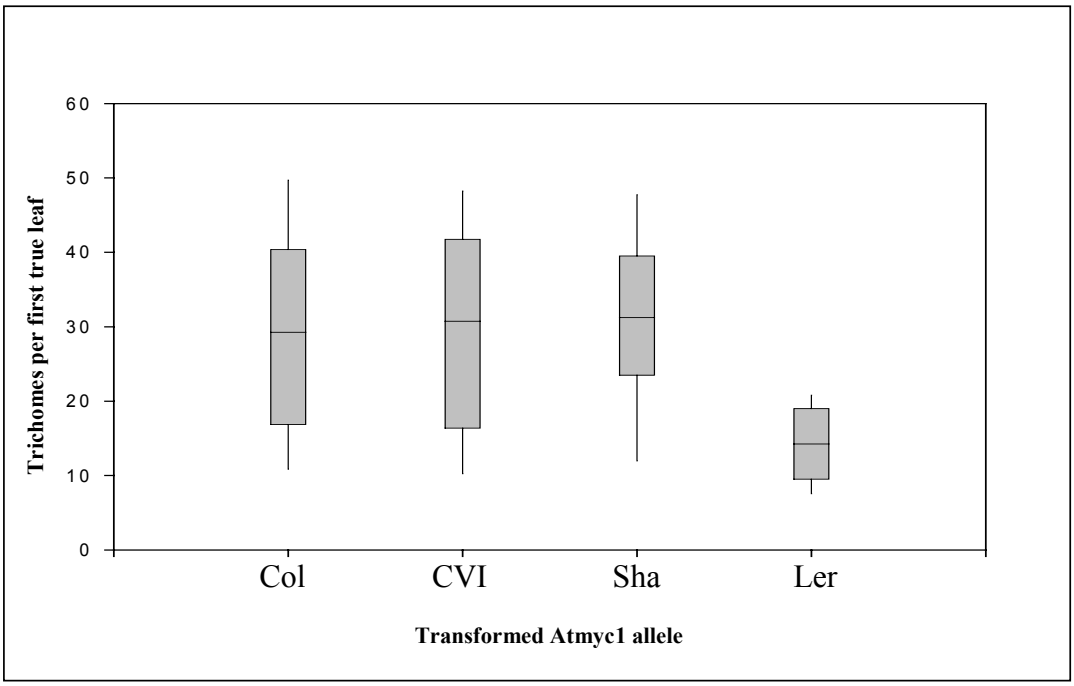

Figure 4.3 Box plots of transformant phenotypes. Each plot represents trichome number data from a pool of *atmyc1* mutant individuals transformed with either the Col, CVI, Sha, or L*er ATMYC1* allele. The median value of transformants for a particular allele is shown as a horizontal bar within the box, the 25th and 75th percentiles are indicated by the top and bottom of each box, and the 10th and 90th percentiles are indicated by vertical line endpoints.

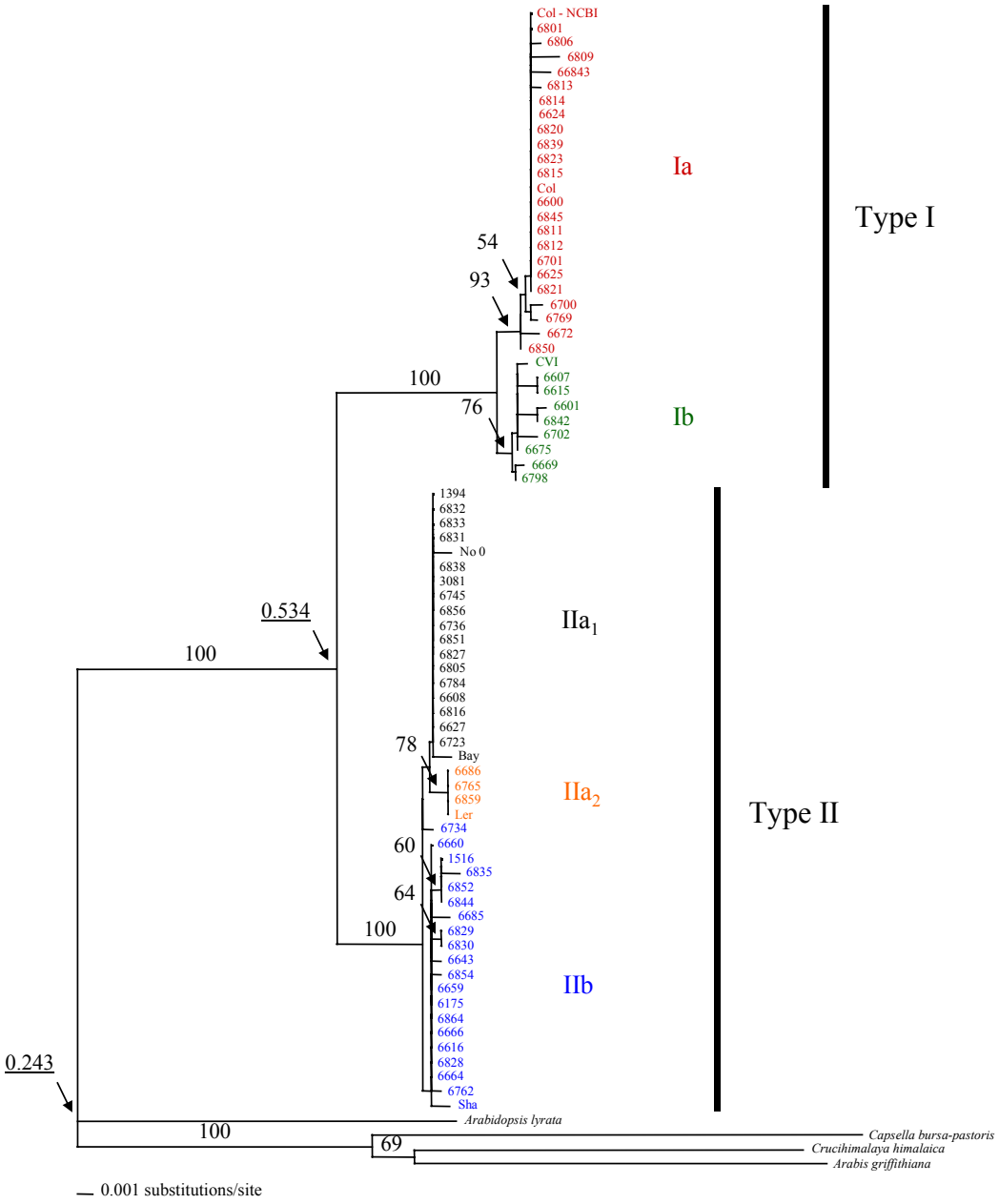

Figure 4.4 Neighbor-joining tree of Amyc1 alleles. Bootstrap values ≥60 are shown above branches. Underlined numbers indicate K_A/K_S ratios for the indicated nodes. Type I, Type II, and subtype allele classifications are indicated at right.

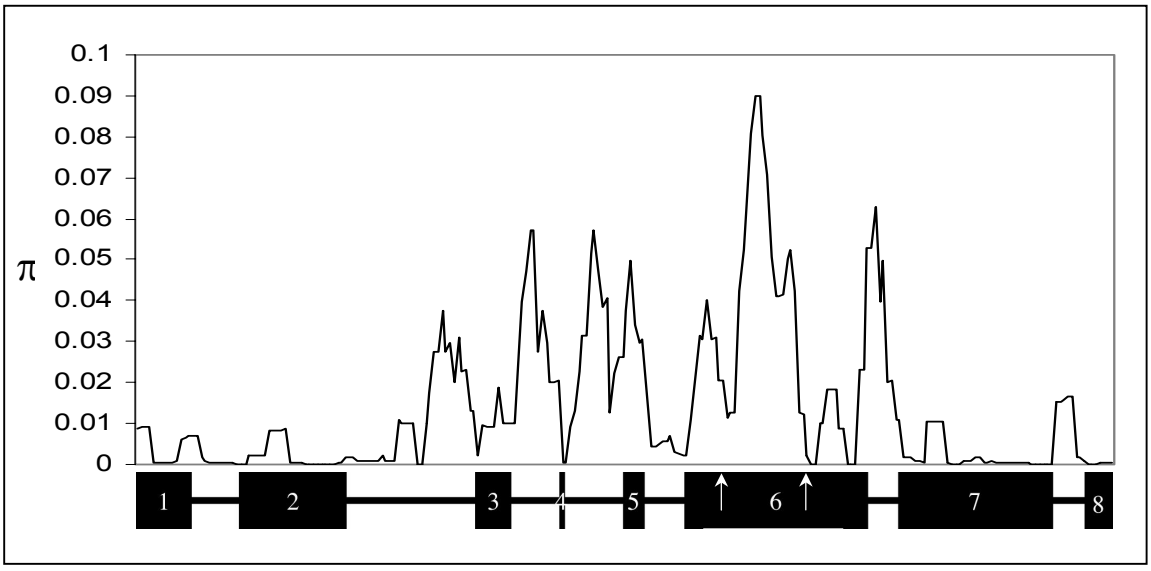

Figure 4.5 Sliding window analysis of nucleotide diversity (π) for the *ATMYC1* alleles of 76 accessions of *A. thaliana* shown aligned with a schematic of the *ATMYC1* locus. Exons are shown as blocks and introns as horizontal lines. Arrows indicate the locations of outgroup indels. Window size was set at 50 bp and moved in 10 bp increments.

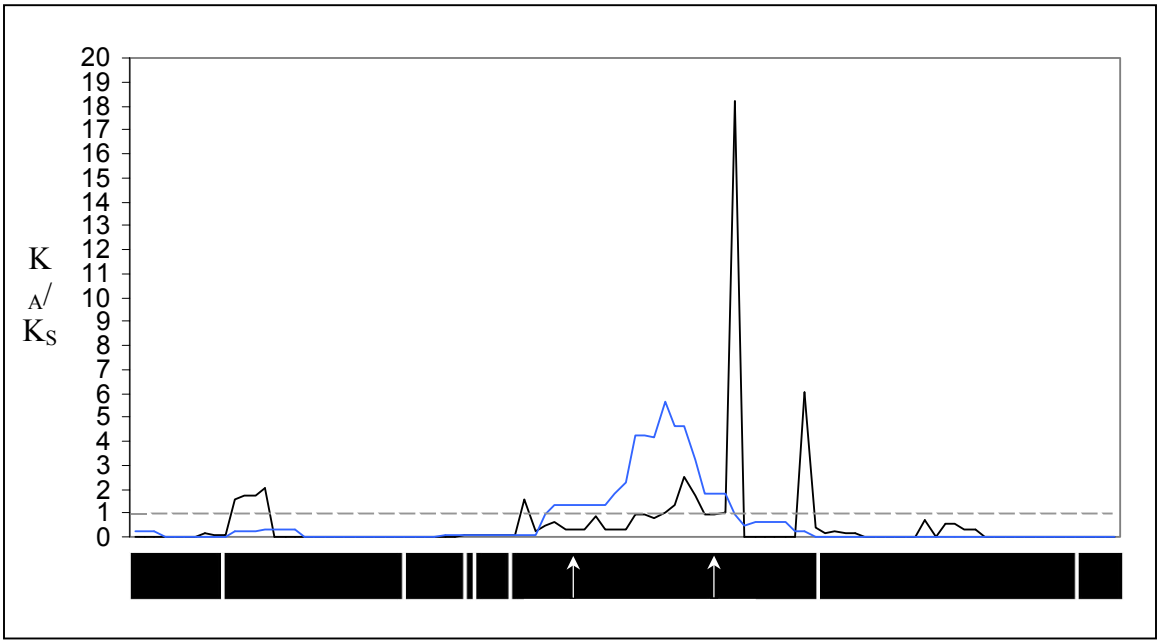

Figure 4.6 Sliding window analysis of the ratio of nonsysnonymous substitution rate (K_A) to synonymous substitution rate (K_S). These plots were generated by comparing two pools of *ATMYC1* alleles (Type I and Type II). The black line represents the ratio of local K_A/local K_S and the blue line represents the ratio of local K_A/gene-wide average K_S. Values near one represent neutral molecular evolution, values <<1 indicate purifying selection and values >>1 indicate positive selection. Plots are aligned with the eight exons of *ATMYC1*, distinguished by vertical white lines. Arrows indicate the locations of outgroup indels.

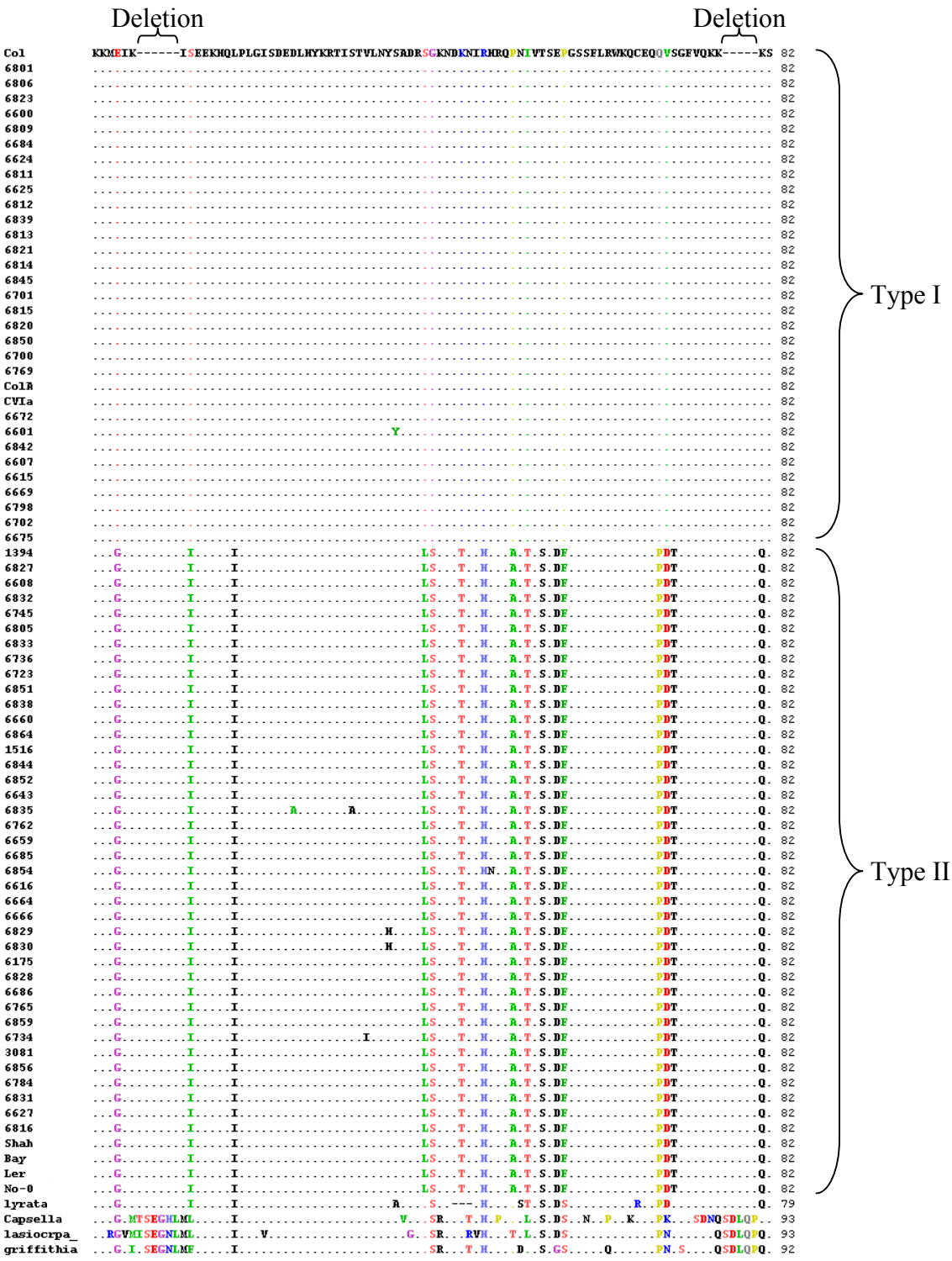

Figure 4.7

Figure 4.7 Deletion block alignment of *ATMYC1* alleles for 76 *A. thaliana* accessions and four outgroup taxa. Accession numbers and taxon names are at left. Types I and II alleles are indicated in clustered blocks and the positions of the "outgroup indels" are marked above. Dots represent identical amino acids and substitutions are shown in color.

References

ABIOLA, O., J. M. ANGEL, P. AVNER, A. A. BACHMANOV, J. K. BELKNAP et al., 2003 The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet 4: 911-6.

ALONSO-BLANCO, C., and M. KOORNNEEF, 2000 Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. Trends in Plant Science 5: 22-29.

ASHMAN, T. L., 2003 Constraints on the evolution of males and sexual dimorphism: field estimates of genetic architecture of reproductive traits in three populations of gynodioecious Fragaria virginiana. Evolution 57: 2012-2025.

BACHTROG, D., M. AGIS, M. IMHOF and C. SCHLÖTTERER, 2000 Microsatellite variability differs between dinucleotide repeat motifs-evidence from Drosophila melanogaster. Mol Biol Evol 17: 1277-85.

BALLOUX, F., and N. LUGON-MOULIN, 2002 The estimation of population differentiation with microsatellite markers. Mol Ecol 11: 155-65.

BARKER, G. C., 2002 Microsatellite DNA: a tool for population genetic analysis. Trans R Soc Trop Med Hyg 96: S21-4.

BARTON, N. H., and P. D. KEIGHTLEY, 2002 Understanding quantitative genetic variation. Nature Reviews 3: 11-21.

BELL, C. J., and J. R. ECKER, 1994 Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics 19: 137-44.

BERGELSON, J., E. STAHL, S. DUDEK and M. KREITMAN, 1998 Genetic variation within and among populations of Arabidopsis thaliana. Genetics 148: 1311-23.

BOHONAK, A. J., 1999 Dispersal, gene flow, and population structure. Quarterly review of biology 74: 21-45.

BOREVITZ, J. O., D. LIANG, D. PLOUFFE, H. S. CHANG, T. ZHU et al., 2003 Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13: 513-23.

BOUTIN-GANACHE, I., M. RAPOSO, M. RAYMOND and C. F. SDESCHEPPER, 2001 M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. Biotechniques 31: 24-28.

BROHEDE, J., C. R. PRIMMER, A. MOLLER and H. ELLEGREN, 2002 Heterogeneity in the rate and pattern of germline mutation at individual microsatellite loci. Nucleic Acids Res 30: 1997-2003.

CASACUBERTA, E., P. PUIGDOMENECH and A. MONFORT, 2000 Distribution of microsatellites in relation to coding sequences within the Arabidopsis thaliana genome. Plant Science 157: 97-104.

CHAKRABORTY, R., M. KIMMEL, D. N. STIVERS, L. J. DAVISON and R. DEKA, 1997 Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A 94: 1041-6.

CLAUSS, M. J., H. COBBAN and T. MITCHELL-OLDS, 2002 Cross-species microsatellite markers for elucidating population genetic structure in Arabidopsis and Arabis (Brassicaceae). Mol Ecol 11: 591-601.

COPENHAVER, G. P., E. A. HOUSWORTH and F. W. STAHL, 2002 Crossover interference in Arabidopsis. Genetics 160: 1631-9.

CORNUET, J. M., and G. LUIKART, 1996 Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics 144: 2001-14.

CRUZAN, M., 1998 Genetic markers in plant evolutionary ecology. Ecology 79: 400-412.

CUNNIFF, C., 2001 Molecular mechanisms in neurologic disorders. Semin Pediatr Neurol 8: 128-34.

DONOHUE, K., E. H. PYLE, D. MESSIQUA, M. S. HESCHEL and J. SCHMITT, 2000 Density dependence and population differentiation of genetic architecrure in Impatiens capensis in natural environments. Evolution 54: 1969-1981.

DOYLE, J. F., and J. L. DOYLE, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf material. Phytochemical Bulletin 19: 11-15.

DRISCOLL, C. A., M. MENOTTI-RAYMOND, G. NELSON, D. GOLDSTEIN and S. J. O'BRIEN, 2002 Genomic microsatellites as evolutionary chronometers: a test in wild cats. Genome Res 12: 414-23.

ECKERT, K. A., A. MOWERY and S. E. HILE, 2002 Misalignment-mediated DNA polymerase beta mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry 41: 10490-8.

EL-DIN EL-ASSAL, S., C. ALONSO-BLANCO, A. J. PEETERS, V. RAZ and M. KOORNNEEF, 2001 A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. Nat Genet 29: 435-40.

ELLEGREN, H., 2000 Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet 16: 551-8.

ERICKSON, D. L., C. B. FENSTER, H. K. STENOIEN and D. PRICE, 2004 Quantitative trait locus analyses and the study of evolutionary process. Mol Ecol 13: 2505-22.

ESTOUP, A., and B. ANGERS, 1998 Microsatellites and minisatellites for molecular ecology: Theoretical and Empirical considerations, pp. 55-86 in Advances in molecular ecology, edited by C. G. R. Nato Sciences Series, IOS Press.

ESTOUP, A., and J. CORNUET, 1999 Microsatellite evolution: inferences from population data, pp. 49-65 in Microsatellites: evolution and applications, edited by G. D. B and S. C. Oxford University Press, New York.

ESTOUP, A., P. JARNE and J. M. CORNUET, 2002 Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11: 1591-604.

FALCONER, D. S., and T. F. MACKAY, 1996 Introduction to Quantitative Genetics. Prentice Hall, Malaysia.

FARRIS, J. S., M. KELLERSJO, A. G. KLUGE and C. BULT, 1994 Testing significance of congruence. Cladistics 10: 315-320.

FARRIS, J. S., M. KELLERSJO, A. G. KLUGE and C. BULT, 1995 Constructing a significance test for incongruence. systematic botany 44: 570-572.

FRARY, A., T. C. NESBITT, S. GRANDILLO, E. KNAAP, B. CONG et al., 2000 fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 289: 85-8.

GILBERT, L., 1971 Butterfly-plant coevolution: has Passiflora adenopoda won the selectional race with Heliconiine butterflies? Science 172: 585-586.

GILL, P., A. J. JEFFREYS and D. J. WERRETT, 1985 Forensic application of DNA 'fingerprints'. Nature 318: 577-9.

GLAZIER, A. M., J. H. NADEAU and T. J. AITMAN, 2002 Finding genes that underlie complex traits. Science 298: 2345-2349.

GRIMALDI, M. C., and B. CROUAU-ROY, 1997 Microsatellite allelic homoplasy due to variable flanking sequences. J Mol Evol 44: 336-40.

GURGANUS, M. C., S. V. NUZHDIN, J. W. LEIPS, and T. F. MACKAY, 1999 High-resolution mapping of quantitative trait loci for sternopleural bristle number in Drosophila melanogaster. Genetics 152: 1585-1604.

HALL, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series 41: 95-98.

HAMRICK, J. L., and M. J. W. GODT, 1996 Effects of life history traits on genetic diversity in plant species. Philos Trans R Soc Lond B Biol Sci 351: 1291-1298.

HANCOCK, J. M., 1999 Microsatellites and other simple sequences: genomic context and mutational mechaisms, pp. 1-9 in Microsatellites: Evolution and Applications, edited by D. Goldstein and C. Schlötterer. Oxford University Press, Inc., New York.

HAUSER, M. T., B. HARR and C. SCHLÖTTERER, 2001 Trichome distribution in Arabidopsis thaliana and its close relative Arabidopsis lyrata: molecular analysis of the candidate gene GLABROUS1. Mol Biol Evol 18: 1754-63.

HEIM, M. A., M. JAKOBY, M. WERBER, C. MARTIN, B. WEISSHAAR et al., 2003 The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol Biol Evol 20: 735-47. Epub 2003 Apr 02.

HILE, S. E., G. YAN and K. A. ECKERT, 2000 Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res 60: 1698-703.

HORIKAWA, Y., N. ODA, N. J. COX, X. LI, M. ORHO-MELANDER et al., 2000 Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus. Nat Genet 26: 163-75.

HSIA, C. C., and W. MCGINNIS, 2003 Evolution of transcription factor function. Curr Opin Genet Dev 13: 199-206.

HULSKAMP, M., S. MISRA and G. JURGENS, 1994 Genetic dissection of trichome cell development in Arabidopsis. Cell 76: 555-66.

INNAN, H., F. TAJIMA, R. TERAUCHI and N. T. MIYASHITA, 1996 Intragenic recombination in the Adh locus of the wild plant Arabidopsis thaliana. Genetics 143: 1761-70.

INNAN, H., R. TERAUCHI and N. T. MIYASHITA, 1997 Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana. Genetics 146: 1441-52.

JUENGER, T., S. SEN, K. STOWE and E. SIMMS, 2004 Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in Arabidopsis thaliana. Genetica in press.

KAMIYA, T., A. KAWABE and N. T. MIYASHITA, 2002 Nucleotide polymorphism at the Atmyb2 locus of the wild plant Arabidopsis thaliana. Genet Res 80: 89-98.

KARABOURNIOTIS, G., and Y. MANETAS, 1995 Trichome density and its protective potential against ultraviolet-B radiation damage during leaf development. Canadian Journal of botany 73: 376-383.

KATTI, M. V., P. K. RANJEKAR and V. S. GUPTA, 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 18: 1161-7.

KAWABE, A., H. INNAN, R. TERAUCHI and N. T. MIYASHITA, 1997 Nucleotide polymorphism in the acidic chitinase locus (ChiA) region of the wild plant Arabidopsis thaliana. Mol Biol Evol 14: 1303-15.

KENNEDY, G. G., 2003 Tomato, pests, parasitoids, and predators: tritrophic interactions involving the genus Lycopersicon. Annu Rev Entomol 48: 51-72. Epub 2001 Sep 28.

KIMURA, M., 1977 Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267: 275-6.

KING, M., and A. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107-116.

KOORNNEEF, M., 1981 The complex syndrome of ttg mutants. Arabidopsis information service 18: 45-51.

KOORNNEEF, M., W. DELLAERT and J. VAN DER VEEN, 1982 EMS- and radiation-induced mutation frequencies at individual loci in Arabidopsis thaliana (L.). Mutation research 93: 109-123.

KRUGLYAK, S., R. T. DURRETT, M. D. SCHUG and C. F. AQUADRO, 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95: 10774-8.

KUBO, S., Y. FUJITA, Y. YOSHIDA, K. KANGAWA, I. TOKUNAGA et al., 2002 Personal identification from skeletal remain by D1S80, HLA DQA1, TH01 and polymarker analysis. J Med Invest 49: 83-6.

LARKIN, J. C., M. L. BROWN and J. SCHIEFELBEIN, 2003 How do cells know what they want to be when they grow up? Lessons from epidermal patterning in Arabidopsis. Annu Rev Plant Biol 54: 403-30.

LARKIN, J. C., N. YOUNG, M. PRIGGE and M. D. MARKS, 1996 The control of trichome spacing and number in Arabidopsis. Development 122: 997-1005.

LEVIN, D., 1973 The role of trichomes in plant defense. The quarterly review of biology 48: 3-15.

LIAKOURA, V., M. STEFANOU, Y. MANETAS, C. CHOLEVAS and G. KARABOURNIOTIS, 1997 Trichome density and its UV-B protective potential are affected by shading and leaf position on the canopy. Environmental and Experimental Botany 38: 223-229.

LISTER, C., and C. DEAN, 1993 Recombinant inbred lines for mapping RFLP and phenotypic markers in Arabidopsis thaliana. The Plant Journal 4: 745-750.

LITT, M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet 44: 397-401.

LONG, A. D., S. L. MULLANEY, L. A. REID, J. D. FRY, C. H. LANGLEY, and T. F. MACKAY, 1995 High resolution mapping of genetic factors affecting abdominal bristle number in Drosophila melanogaster. Genetics 139: 1273-1291.

LOUDET, O., S. CHAILLOU, C. CAMILLERI and D. BOUCHEZ, 2002 Bay-0 x Shahdara recombinant inbred line populations: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theoretical and Applied Genetics 104: 1173-1184.

LUKOWITZ, W., C. S. GILLMORE and W. SCHEIBLE, 2000 Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you. Plant Physiology 123: 795-805.

LYNCH, M., and B. WALSH, 1998 Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Sunderland, MA.

MACKAY, T. F., 2001 The genetic architecture of quantitative traits. Annu Rev Genet 35: 303-39.

MATSUOKA, Y., S. E. MITCHELL, S. DRESOVICH, M. GOODMAN and J. DOEBLEY, 2002a Microsatellites in Zea - variability , patterns of mutations, and use for evolutionary studies. Theor Appl Genet 104: 436-450.

MAURICIO, R., 1998 Costs of resistance to natural enemies in field populations of the annual plant Arabidopsis thaliana. American Naturalist 151: 20-28.

MCCOUCH, S. R., X. CHEN, O. PANAUD, S. TEMNYKH, Y. XU et al., 1997 Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35: 89-99.

METZGAR, D., E. THOMAS, C. DAVIS, D. FIELD and C. WILLS, 2001 The microsatellites of Escherichia coli: rapidly evolving repetitive DNAs in a non-pathogenic prokaryote. Mol Microbiol 39: 183-90.

MORGAN, K. K., J. HICKS, K. SPITZE, L. LATTA, M. E. PFRENDER et al., 2001 Patterns of genetic architecture for life-history traits and molecular markers in a subdivided species. Evolution 55: 1753-1761.

MORIGUCHI, Y., H. IWATA, T. UJINO-IHARA, K. YOSHIMURA, H. TAIRA et al., 2003 Development and characterization of microsatellite markers for Cryptomeria japonica D.Don. Theor Appl Genet 106: 751-8.

NEI, M., 1973 Analysis of gene diversity in subdivided populations. PNAS 70: 3321-3323.

NEI, M., 1987 Molecular Evolutionary Genetics. Columbia University Press, New York, NY.

NOEL, L., T. L. MOORES, E. A. VAN DER BIEZEN, M. PARNISKE, M. J. DANIELS et al., 1999 Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of Arabidopsis. Plant Cell 11: 2099-112.

NOOR, M. A., R. M. KLIMAN and C. A. MACHADO, 2001 Evolutionary history of microsatellites in the obscura group of Drosophila. Mol Biol Evol 18: 551-6.

PARAN, I., and D. ZAMIR, 2003 Quantitative traits in plants: beyond the QTL. Trends Genet 19: 303-6.

PAYNE, C. T., F. ZHANG and A. M. LLOYD, 2000 GL3 encodes a bHLH protein that regulates trichome development in arabidopsis through interaction with GL1 and TTG1. Genetics 156: 1349-62.

POOS, F., 1929 Leafhopper injury to legumes. Journal of economic entomology 22: 146-153.

PRESGRAVES, D. C., L. BALAGOPALAN, S. M. ABMAYR and H. A. ORR, 2003 Adaptive evolution drives divergence of a hybrid inviability gene between two species of Drosophila. Nature 423: 715-9.

RANUM, L. P., and J. W. DAY, 2002 Dominantly inherited, non-coding microsatellite expansion disorders. Curr Opin Genet Dev 12: 266-71.

RICHARD, G. F., and F. PAQUES, 2000 Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1: 122-6.

RIESEBERG, L. H., A. WIDMER, A. M. ARNTZ and J. M. BURKE, 2003 The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations. Philos Trans R Soc Lond B Biol Sci 358: 1141-7.

ROBIN, C., R. F. LYMAN, A. D. LONG, C. H. LANGLEY and T. F. MACKAY, 2002 hairy: A quantitative trait locus for drosophila sensory bristle number. Genetics 162: 155-64.

ROLFSMEIER, M. L., and R. S. LAHUE, 2000 Stabilizing effects of interruptions on trinucleotide repeat expansions in Saccharomyces cerevisiae. Mol Cell Biol 20: 173-80.

ROZAS, J., J. C. SANCHEZ-DEL BARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

RUBINSZTEIN, D. C., B. AMOS and G. COOPER, 1999 Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos Trans R Soc Lond B Biol Sci 354: 1095-9.

SAKAMOTO, T., and N. OKAMOTO, 2000 [Microsatellite linkage map of rainbow trout and its application for QTL analysis]. Tanpakushitsu Kakusan Koso 45: 2872-9.

SANCHEZ-MORAN, E., S. J. ARMSTRONG, J. L. SANTOS, F. C. FRANKLIN and G. H. JONES, 2002 Variation in chiasma frequency among eight accessions of Arabidopsis thaliana. Genetics 162: 1415-22.

SANDQUIST, D., and J. EHLERINGER, 2003 Population- and family-level variation of Brittlebush (Encelia farinosa, Asteraceae) pubescence: its relation to drought and implications for selection in variable environments. American Journal of Botany 90: 1481-1486.

SCHUG, M. D., K. A. WETTERSTRAND, M. S. GAUDETTE, R. H. LIM, C. M. HUTTER et al., 1998 The distribution and frequency of microsatellite loci in Drosophila melanogaster. Mol Ecol 7: 57-70.

SEN, S., and G. CHURCHILL, 2001 A statistical framework for quantitative trait mapping. Genetics 159: 371-387.

SHARBEL, T. F., B. HAUBOLD and T. MITCHELL-OLDS, 2000 Genetic isolation by distance in Arabidopsis thaliana: biogeography and postglacial colonization of Europe. Mol Ecol 9: 2109-18.

SHIU, S. H., W. M. KARLOWSKI, R. PAN, Y. H. TZENG, K. F. MAYER et al., 2004 Comparative analysis of the receptor-like kinase family in Arabidopsis and rice. Plant Cell 16: 1220-34. Epub 2004 Apr 22.

SHRIVER, M. D., J. L., R. CHAKRABORTY and L. E. BOERWINKLE, 1993 VNTR allele frequency distribution under the stepwise mutation model. Genetics 134: 983-993.

SIA, E. A., R. J. KOKOSKA, M. DOMINSKA, P. GREENWELL and T. D. PETES, 1997 Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol 17: 2851-8.

SMITH, W., and P. NOBEL, 1977 Influences of seasonal changes in leaf morphology on water-use efficiency for three desert broadleaf shrubs. Ecology 58: 1033-1043.

SOKAL, R. R., and F. J. ROHLF, 1995 Biometry. W. H. Freeman and Company, New York.

SWOFFORD, D. L., 2002 PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA.

SYMONDS, V. V., A. V. GODOY, T. ALCONADA, J. F. BOTTO, T. J. JUENGER et al., in press Mapping quantitative trait loci in multiple populations of Arabidopsis thaliana identifies natural allelic variation for trichome density. Genetics.

SZYMANSKI, D. B., A. M. LLOYD and M. D. MARKS, 2000 Progress in the molecular genetic analysis of trichome initiation and morphogenesis in Arabidopsis. Trends Plant Sci 5: 214-9.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN and D. G. HIGGINS, 1997 The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic acids research 25: 4876-4882.

TIAN, D., H. ARAKI, E. STAHL, J. BERGELSON and M. KREITMAN, 2002 Signature of balancing selection in Arabidopsis. Proc Natl Acad Sci U S A 99: 11525-30. Epub 2002 Aug 09.

TOLEDO-ORTIZ, G., E. HUQ and P. H. QUAIL, 2003 The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 15: 1749-70.

TORII, K. U., N. MITSUKAWA, T. OOSUMI, Y. MATSUURA, R. YOKOYAMA et al., 1996 The Arabidopsis ERECTA gene encodes a putative receptor protein kinase with extracellular leucine-rich repeats. Plant Cell 8: 735-46.

THUILLET, A. C., D. BRU, J. DAVID, P. ROUMET, S. SANTONI et al., 2002 Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, Triticum turgidum (L.) Thell. ssp durum desf. Mol Biol Evol 19: 122-5.

URAO, T., K. YAMAGUCHI-SHINOZAKI, N. MITSUKAWA, D. SHIBATA and K. SHINOZAKI, 1996 Molecular cloning and characterization of a gene that encodes a MYC-related protein in Arabidopsis. Plant Mol Biol 32: 571-6.

VAN OOIJEN, J., M. BOER, R. JANSEN and C. MALIEPAARD, 2002 MapQTL 4.0, software for the calculation of QTL positions on genetic maps, pp. Plant Research International, Wageningen, the Netherlands.

VAN OOIJEN, J., and R. VOORRIPS, 2001 JoinMap 3.0, Software for the calculation of genetic linkage maps, pp. Plant Research International, Wageningen, the Netherlands.

VAN TREUREN, R., H. KUITTINEN, K. KARKKAINEN, E. BAENA-GONZALEZ and O. SAVOLAINEN, 1997 Evolution of microsatellites in Arabis petraea and Arabis lyrata, outcrossing relatives of Arabidopsis thaliana. Mol Biol Evol 14: 220-9.

VIGOUROUX, Y., J. S. JAQUETH, Y. MATSUOKA, O. S. SMITH, W. D. BEAVIS et al., 2002 Rate and pattern of mutation at microsatellite loci in maize. Mol Biol Evol 19: 1251-60.

VISION, T. J., D. G. BROWN and S. D. TANKSLEY, 2000 The origins of genomic duplications in Arabidopsis. Science 290: 2114-7.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theoretical population biology 7: 256-276.

WIERDL, M., M. DOMINSKA and T. D. PETES, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics 146: 769-79.

ZANE, L., L. BARGELLONI and T. PATARNELLO, 2002 Strategies for microsatellite isolation: a review. Mol Ecol 11: 1-16.

ZHANG, F., A. Gonzalez, M. Zhao, C. T. Payne and A. Lloyd, 2003 A network of redundant bHLH proteins functions in all TTG1-dependent pathways of Arabidopsis. Development 130: 4859-69. Epub 2003 Aug 13.

ZIMMERMANN, I. M., M. A. HEIM, B. WEISSHAAR and J. F. UHRIG, 2004 Comprehensive identification of Arabidopsis thaliana MYB transcription factors interacting with R/B-like BHLH proteins. Plant J 40: 22-34.

ZWETTLER, D., C. P. VIEIRA and C. SCHLÖTTERER, 2002 Polymorphic microsatellites in Antirrhinum (Scrophulariaceae), a genus with low levels of nuclear sequence variability. The Jounal of Heredity 93: 217-221.

Vita

Victor Vaughan Symonds II was born in Honolulu, HI on June 18, 1972, the son of Victor Vaughan Symonds I and Betty Joan Symonds. Following graduation from Monache High School in Porterville, CA in 1990, he attended Porterville Jr. College and Fresno City College. In 1993, he was accepted to Fresno State University. After completing his Bachelor of Arts degree in Environmental Biology in 1995, he remained at Fresno State to earn his Master of Arts degree in Biology. This degree program was completed in 1998. In September of 1998, he entered the Ph.D. program at the University of Texas at Austin.

Permanent address: 2511 Enfield Rd., Apt. D, Austin, TX 78703

This dissertation was typed by the author.